Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering

Costa Ballena, Rota, Cádiz (Spain)

July 3rd-7th, 2014

CMMSE 2014

VOLUME I

Editor: J. Vigo-Aguiar

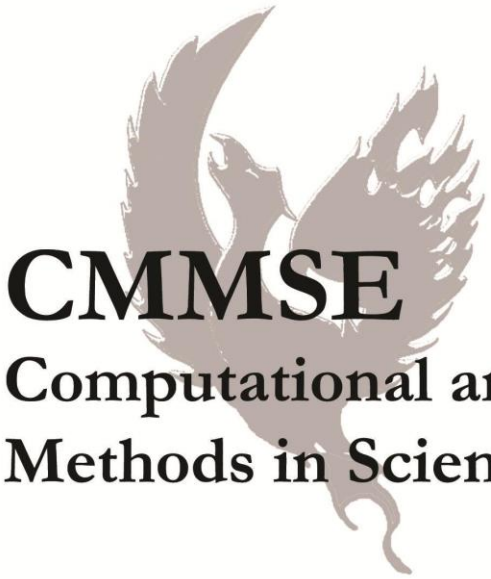
Associate Editors

I. P. Hamilton, J. Medina, P. Schwerdtfeger, W. Sprößig,
M. Demiralp, E. Venturino, V.V. Kozlov, P. Oliveira

Proceedings of the 2014 International Conference on Computational and Mathematical Methods in Science and Engineering

Costa Ballena (Rota), Cádiz, Spain

July 3-7, 2014



CMMSE

Computational and Mathematical Methods in Science and Engineering

Editor:

J. Vigo-Aguiar

Associate Editors

I. P. Hamilton, J. Medina, P. Schwerdtfeger, W. Sprößig, M. Demiralp, E. Venturino, V.V. Kozlov, P. Oliveira

ISBN **978-84-616-9216-3**

Printed on acid-free paper

Preface

One year more, it is our great pleasure to present the proceedings of the **14th International Conference on Computational and Mathematical Methods in Science and Engineering** (CMMSE 2014), at Rota, Cádiz (Spain), July 3rd-7th, 2014. These proceedings, comprised of the extended abstracts and the papers accepted to the conference, are of significant interest and contain original and substantial analyses of computational and mathematical methodologies. The proceedings have five volumes, the first four correspond to the articles typeset in LaTeX and the fifth to articles typeset in Word.

CMMSE 2014 continues with the same philosophy of being a great forum where researchers from several disciplines of applied mathematics discus about the new advances and open problems. We hope that during the session the usual and desirable exchange of ideas, comments and suggestions leading to the improvement and deepening of the papers presented to allow further development of the research occurs. We also hope that the developed activity narrows and renews the links between participants.

This year we have achieved a new record in the number of symposiums and the quality of the accepted papers is also very high. The first one, *high-performance computing*, considers new large-scale problems that arise in fields like bioinformatics, computational chemistry, and astrophysics. The second symposium address analytical, numerical and computational aspects of *partial differential equations in life and materials science*. *Computational finance* is a session focuses on solving problems related to asset pricing, trading and risk analysis of financial assets that have no analytic solutions under realistic assumptions and thus require computational methods to be resolved. A forum for discussion of the growing impact of new technologies on teaching and the development of new tools to increase learning efficiency is provided in the symposium: *new educational methodologies supported by new technologies offers*. The symposium on *mathematical models and information-intelligent transport systems* researches in the field of flow-modelling of particles with motivated behaviour in complex networks, applied to traffic flows, pedestrian flows, ecology, etc. Special utility has this symposium in the traffic regulation in Moscow. The seventh symposium studies *computational methods for linear and nonlinear optimization* and *numerical methods for solving nonlinear problems* is given in another session. *Bio-mathematics* studies both theoretical and practical applications of population dynamics, eco-epidemiology, epidemiology of infectious diseases and molecular and antifenic evolution. The tenth symposium presents *recent methodological developments in function approximation, multiway array decompositions, ODE and PDE solutions: applications from dynamical systems to quantum and statistical dynamics*. The *applications of fractional derivatives in sciences* are considered in the twelfth symposium. The *applied mathematics and computer science* symposium focuses on new methods, technologies and applications of computer science and mathematics. Fractional Calculus from a theoretical viewpoint is considered in *analytical and numerical methods for fractional differential equations*. Obtaining a consistent description of the transition from small clusters to the liquid or solid state is a major challenge in computational chemistry and physics and will be addressed in the symposium: *from clusters to the solid state*. *Crypto & codes* aims to provide a forum there researchers can exchange the latest results, trends and open problems in the areas of cryptography and coding theory. *Computational methods for fluid flow* uses numerical methods and algorithms to solve and analyze the mathematical models that govern fluid

flows. Various *numerical solution methods for large linear systems* are presented and discussed in this session. The enormous potential of fixed point theory, which is needed in mathematics, engineering, chemistry, biology, economics, computer science, and other sciences, justify the great interest in *fixed point theory in various abstract spaces and related applications*. Finally, special sessions cover topics related to industrial mathematics, computational discrete mathematics and the numerical solution of differential equations.

We would like to thank the plenary speakers for their outstanding contributions to research and leadership in their respective fields, including physics, chemistry and engineering. We would also like to thank the special session organizers and scientific committee members, who have played a very important part in setting the direction of CMMSE 2014. Finally, we would like to thank the participants because, without their interest and enthusiasm, the conference would not have been possible.

We cordially welcome all participants. We hope you enjoy the conference.

Costa Ballena, Rota, Cádiz (Spain), July 15th, 2014

I. P. Hamilton, J. Vigo-Aguiar, J. Medina,
P. Schwerdtfeger, W. Sprößig, M. Demiralp,
E. Venturino, V.V. Kozlov, P. Oliveira

CMMSE 2014 Mini-symposia

Session Title	Organizers
High Performance Computing (HPC)	
P.D.E.'S in Life and Material Sciences	Paula Oliveira & J.A. Ferreira
Computational Finance	Luis Ortiz Gracia & Carlos Vázquez Cendón
New Educational Methodologies Supported by New Technologies	
Mathematical Models and Information-Intelligent Systems on Transport	Valerii V. Kozlov & Andreas Schadschneider & Alexander P. Buslaev
Computational Methods for Linear and Nonlinear Optimization	Maria Teresa Torres Monteiro
Numerical Methods for Solving Nonlinear Problems	Juan R. Torregrosa & A. Cordero
Bio-mathematics	Ezio Venturino, Nico Stollenwerk & Maíra Aguiar
Recent Methodological Developments in Function Approximation, Multiway Array Decompositions, ODE and PDE Solutions: Applications From Dynamical Systems to Quantum and Statistical Dynamics	Metin Demiralp & Alper Tunga & Burcu Tunga
Mathematical Models for Computer Science	Jesús Medina & Manuel Ojeda-Aciego
Analytical and numerical methods for fractional differential equations	Luisa Morgado
From clusters to the solid state	Ian Hamilton & Peter Schwerdtfeger
Computational methods for fluid flow	Zhenquan Li
Fixed Point Theory in various abstract spaces and related applications	Antonio F. Roldán López de Hierro & Juan Martínez Moreno

Acknowledgements

We would like to express our gratitude to Universidad de Cádiz, our sponsor, for its assistance.

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

- Jesús Medina Moreno, Universidad de Cádiz, Spain
- Juan Carlos Díaz Moreno- Universidad de Cádiz, Spain
- María Eugenia Cornejo Piñero - Universidad de Cádiz, Spain
- Eloisa Ramírez Poussa - Universidad de Cádiz, Spain

CMMSE 2014 Plenary Speakers

- Carlos Vázquez Cendón - University of A Coruña, **Spain**
- Peter Schwerdtfeger - University in Auckland, **New Zealand**
- Prof. Dr. Wolfgang Sprößig - TU Bergakademie Freiberg **Germany**
- Maira Aguiar - University Lisbon, **Portugal**
- Metin Demiralp - Istanbul Technical University, **Turkish**
- Ezio Venturino - University of Torino, **Italy**
- Valery V. Kozlov - Russia Academy of Sciences, **Russia**
- Paula Oliveira - University of Coimbra, **Portugal**

Contents:

Volume I

Contents:

Volume II

Contents:

Volume III

Contents:

Volume IV

Contents:

Volume V

Volume I

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

The Ikebe-Schulz algorithm for inverting Hessenberg matrices

J. Abderramán Marrero[1]

[1] *Department of Mathematics Applied to Information Technologies, Telecommunication Engineering School, UPM Technical University of Madrid, Spain*

emails: jc.abderraman@upm.es

Abstract

Two algorithms are proposed for inverting general nonsingular upper Hessenberg matrices $\mathbf{H}$, unreduced as well as reduced. The first step in both procedures is based on the expanded Ikebe method, now adapted to work also on reduced matrices. Thus the quasiseparable lower Hessenberg matrix $\mathbf{H}_L$ of the inverse factorization $\mathbf{H}^{-1} = \mathbf{H}_L\mathbf{U}^{-1}$ is obtained. Then the inverse is obtained by using two well-known numerically stable procedures. In the first algorithm the matrix $\mathbf{U}^{-1}$ is computed by forward substitution (column version). The second one uses the iterative (Newton's method) Schulz algorithm, with $\mathbf{H}_L$ as initial guess. It is shown that such an iterative procedure provides an optimal route towards the inverse matrix $\mathbf{H}^{-1}$ by increasing the superdiagonal rank of the quasiseparable matrix $\mathbf{H}_L$. An illustrative numerical comparison is also given.

Key words: Accuracy, Hessenberg matrix, inverse matrix, iterative Newton's method, matrix factorization, stability.

1.1 Introduction

The inverses matrices $\mathbf{H}^{-1}$ of $n \times n$ nonsingular Hessenberg matrices $\mathbf{H}$, with entries $h_{i,j}$ $(1 \leq i, j \leq n)$, $h_{i,j} = 0$ for $i - j \geq 2$ in the upper case treated here, appear in many branches of mathematics, applied sciences and engineering. Abundant literature has been generated in the particular study of such inverses; see e.g. [3, 1] and references given there. These approaches have been mainly focused on unreduced Hessenberg matrices, with nonzero entries on their subdiagonals. A new factorization $\mathbf{H}^{-1} = \mathbf{H}_L\mathbf{U}^{-1}$ for the inverses of

 ISBN: 978-84-616-9216-3

nonsingular upper Hessenberg matrices and its related algorithm in the unreduced case was introduced recently in [1], which expands the Ikebe method [3] to obtain the quasiseparable matrix $\mathbf{H}_L$ and apply forward substitution to obtain the triangular matrix $\mathbf{U}^{-1}$ involved in such an inverse factorization. Specific methods for inverting reduced Hessenberg matrices are also of interest, different to those based on the Schur complement approach, which presents in general bad numerical performance [2].

Now we adapt the expand Ikebe method to cover nonsingular reduced Hessenberg matrices by partitioning the quasiseparable matrix $\mathbf{H}_L$ as a block diagonal matrix. Hence the algorithm from [1] is now applicable also on reduced Hessenberg matrices. In addition, we can improve the accuracy, with respect to the forward substitution step for obtaining $\mathbf{U}^{-1}$ given in the algorithm from [1], by increasing the superdiagonal rank of $\mathbf{H}_L$ up to complete the whole inverse $\mathbf{H}^{-1}$. It is proven that this procedure is nothing else but the iterative Schulz algorithm (Newton's method) [4], with the matrix $\mathbf{H}_L$ as initial guess. Such an initial election enables us to attain, in almost all situations, the numerical inverse of an $n \times n$ nonsingular upper Hessenberg matrix over $ceil(\log_2(n))$ iterations. This number is in practice notably less than the number of iterations required for the Schulz algorithm with its customary initial guess.

1.2 The Ikebe method on reduced Hessenberg matrices

For upper unreduced Hessenberg matrices $\mathbf{H}$, the Ikebe method to obtain the lower half (plus the main diagonal) of their inverses matrices $\mathbf{H}^{-1}$ in $O(n^2)$ time was introduced in [3]. Recently, the Ikebe method was expanded to cover the superdiagonal of the inverse matrix [1] without additional computational cost.

There is no difficulty to adapt the expanded Ikebe method to reduced Hessenberg matrices noting that such reduced matrices $\mathbf{H}$ can be partitioned in a block upper triangular matrix. Then we can apply the expanded Ikebe algorithm on each matrix entry of the main diagonal and the resulting matrix $\mathbf{H}_L$ is a (block diagonal) quasiseparable lower Hessenberg matrix. The superdiagonal entries in the exterior of the diagonals blocks of the inverse matrix $\mathbf{H}^{-1}$ can be chosen arbitrarily, in the positions opposite to those ($h_{i,i-1} = 0$) null entries on the subdiagonal of the reduced Hessenberg matrix $\mathbf{H}$. It is because in the reduced case any election of such inverse entries preserves the inverse factorization $\mathbf{H}^{-1} = \mathbf{H}_L\mathbf{U}^{-1}$, with the upper triangular matrix $\mathbf{U}^{-1}$ having ones on the main diagonal. We choose, by numerical convenience, such entries of the inverse as null entries. Note also that this inverse factorization is not unique. Although solely one of these has the matrix $\mathbf{H}_L$ with the same entries $(h_{i,j})^{(-1)}$ than $\mathbf{H}^{-1}$ for $j \leq i + 1$.

 ISBN: 978-84-616-9216-3

1.3 Obtaining $\mathbf{U}^{-1}$ by forward substitution

With the expanded Ikebe procedure adapted to cover also reduced Hessenberg matrices, there is no difficulty to apply a forward substitution vector scheme (column version) on the matrix $\mathbf{U} = \mathbf{H}\mathbf{H}_L$ to obtain the matrix $\mathbf{U}^{-1}$ after a slight modification of the algorithms from [1]. See also Equation (13) in the same reference and some comparisons of the times elapsed with respect to other specific algorithm for inverting unreduced Hessenberg matrices.

1.4 The Ikebe-Schulz algorithm

An alternative procedure to the forward substitution stage that improves the accuracy of the numerical inverses is based on the obtainment of the inverse matrix $\mathbf{H}^{-1}$ by completing in an iterative way the diagonals of the upper half (above the main diagonal) of the matrix $\mathbf{H}_L$. It is not difficult to observe that, taking $\mathbf{H}_L^{(0)} = \mathbf{H}_L$ and defining the iterative procedure (Newton's method) $\mathbf{H}_L^{(k+1)} = \mathbf{H}_L^{(k)}\left(2\mathbf{I}_n - \mathbf{H}\mathbf{H}_L^{(k)}\right)$, $k = 0, 1, 2, \ldots$, the matrix $\mathbf{H}_L^{(k+1)}$ updates in each iteration some of the diagonals of its upper half, in such a way that the entries of these diagonals are equal than those of the corresponding diagonals of the inverse matrix $\mathbf{H}^{-1}$. The following theorem gives explanation for such assertions.

Theorem 1 *Let $\mathbf{H}$ be an $n \times n$ nonsingular upper Hessenberg matrix and $\mathbf{H}_L$ the quasiseparable matrix obtained with the expanded Ikebe algorithm and associated to the inverse factorization $\mathbf{H}^{-1} = \mathbf{H}_L\mathbf{U}^{-1}$. The iterate matrices $\mathbf{H}_L^{(k)}$ of Newton's method, with initial guess $\mathbf{H}_L^{(0)} = \mathbf{H}_L$, satisfy the equation, for $k = 0, 1, 2, \ldots$,*

$$\mathbf{H}_L^{(k)} \;=\; \mathbf{H}^{-1}\left(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{U})^{2^k}\right).$$

Moreover, the inverse matrix can be obtained in at most $ceil(\log_2(n))$, where $ceil(x)$ denotes the smaller integer greater or equal than x.

Sketch of proof: Note that $\mathbf{H}_L^{(0)} = \mathbf{H}_L = \mathbf{H}^{-1}\mathbf{U}$. Hence $\mathbf{H}_L^{(0)} = \mathbf{H}^{-1}\left(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{U})\right)$. Thus the equation can be checked by induction. To obtain the inverse matrix we use Cayley-Hamilton's theorem on matrix $\mathbf{I}_n - \mathbf{U}$, $(\mathbf{I}_n - \mathbf{U})^n = \mathbf{O}_n$, the zero square matrix. Thus matrix $\mathbf{I}_n - \mathbf{U}$ is nilpotent. Therefore for $ceil(\log_2(n)) \leq k$, we have $\mathbf{H}_L^{(k)} = \mathbf{H}^{-1}$. □

Theorem 1 gives rise to the Ikebe-Schulz algorithm for inverting Hessenberg matrices. The first stage is the expanded Ikebe algorithm applicable to arbitrary nonsingular Hessenberg matrices. In the second stage we obtain the inverse matrix by using Schulz's algorithm [4], with $\mathbf{H}_L$ as initial guess.

In Figure 1, a comparison is done of the mean left, $\dfrac{||\hat{\mathbf{H}}^{-1}\mathbf{H} - \mathbf{I}_n||_\infty}{||\hat{\mathbf{H}}^{-1}||_\infty||\mathbf{H}||_\infty}$, and the mean right normwise relative residual [2], $\dfrac{||\mathbf{H}\hat{\mathbf{H}}^{-1} - \mathbf{I}_n||_\infty}{||\mathbf{H}||_\infty||\hat{\mathbf{H}}^{-1}||_\infty}$, over 50 trials, where $\hat{\mathbf{H}}^{-1}$ is the numerical

 ISBN: 978-84-616-9216-3

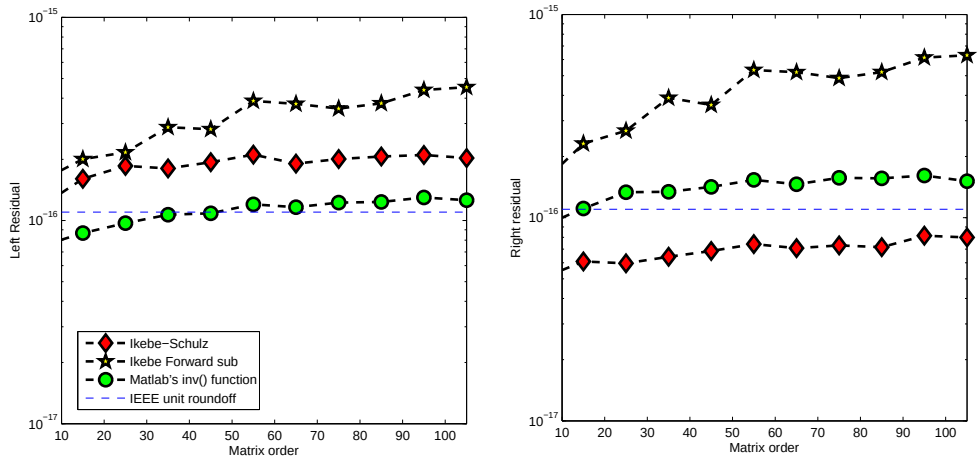


Figure 1: Comparison of the accuracy using mean relative residuals, over 50 trials, for the three procedures used for computing the inverses of random nonsingular upper Hessenberg matrices $\mathbf{H}$, with $rcond(\mathbf{H}) > 10^{-3}$.

inverse, supplied by the algorithms Ikebe-Schulz, Ikebe-forward substitution, and *Matlab's*® build-in function *inv*(). Random unreduced upper Hessenberg matrices $\mathbf{H}$ were tested with reciprocal condition numbers $rcond(\mathbf{H}) > 10^{-3}$. The order of matrices from 15 to 105, in step of 10 units. The IEEE unit roundoff, $u \approx 1.1 \cdot 10^{-16}$, is also indicated. The accuracy of the right inverse of the Ikebe-Schulz algorithm is remarkable.

References

[1] J. Abderramán Marrero, M. Rachidi, V. Tomeo, *On new algorithms for inverting Hessenberg matrices*, J. Comp. Appl. Math. **252** (2013) 12–20.

[2] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, U.S.A., 1996.

[3] Y. Ikebe, *On inverses of Hessenberg matrices*, Linear Algebra Appl. **24** (1979) 93–97.

[4] G. Schulz, *Iterative berechnung der reziproken matrix*, Z. Angew. Math. Mech. **13** (1933) 57–59.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Inversion of infinite tridiagonal matrices

J. Abderramán Marrero[1], Venancio Tomeo[2] and Emilio Torrano[3]

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

[2] *Department of Algebra, Faculty of Statistical Studies, University Complutense of Madrid, Spain*

[3] *Department of Applied Mathematics, Faculty of Informatics, UPM - Technical University of Madrid, Spain*

emails: jc.abderraman@upm.es, tomeo@estad.ucm.es, emilio@fi.upm.es

Abstract

A method, based on recurrence relations, is proposed for evaluating classical inverses of infinite (unreduced as well as reduced) tridiagonal matrices using a known result on inversion of finite Hessenberg matrices, applicable also on tridiagonal matrices. Some illustrative examples of both unreduced and reduced cases are given. The recurrences relations for the inverse way are also provided.

Key words: Hessenberg matrix, tridiagonal matrix, inverse matrix.

1 Introduction

A characterization for the nonsingular unreduced Hessenberg matrices in the finite case is related with the particular structure of their inverse matrices, [5, 6]. Such inverses are a rank one perturbation of a triangular matrix $UV + T$. Matrix T is triangular, U is a column vector and V is a row vector. This result can be also applied to the case of finite tridiagonal matrices. In this work we study the case of infinite tridiagonal matrices and we find recurrence relations for their classical inverses in the general case. Some interesting particular cases are also discussed. In this work we do not view infinite matrices as operators on some vector spaces, only as matrices on $\mathbb{R}$ or $\mathbb{C}$.

The material is organized as follow. In Section 1, we recall some basic result about inverses of finite Hessenberg and tridiagonal matrices. In Section 2 we study the inversion

of infinite unreduced tridiagonal matrices and, as a corollary, we give the inverse way. Finally, Section 3 covers the case of infinite reduced tridiagonal matrices. Some illustrative examples are also given.

1.1 Unreduced Hessenberg matrices with a finite order

We extend and adapt here to upper Hessenberg matrices H; i.e. $h_{ij} = 0$ for $i \geq j + 2$, a well-known lemma [5, 6]. We also recall that a Hessenberg matrix $H = (h_{ij})_{i,j=1}^{n}$ is an unreduced upper Hessenberg matrix if its subdiagonal entries are nonzero, $h_{i+1,i} \neq 0$, and $i = 1, 2, ..., n - 1$.

Lemma 1 *A nonsingular matrix $H = (h_{ij})_{i,j=1}^{n}$ is unreduced upper Hessenberg if and only if its inverse matrix has the structure $B = UV + T$, being U a column matrix with nonzero n-th component, V is a row matrix with nonzero 1-st component, and T is a strictly upper triangular having null entries on its main diagonal and nonzero entries on the superdiagonal, $t_{i,i+1} = \frac{1}{h_{i+1,i}} \neq 0, 1 \leq i \leq n - 1$.*

In the proof of Lemma 1 appears the matrix

$$B = \begin{pmatrix} u_1v_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ u_2v_1 & u_2v_2 & b_{23} & \cdots & b_{2n} \\ u_3v_1 & u_3v_2 & u_3v_3 & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_nv_1 & u_nv_2 & u_nv_3 & \cdots & u_nv_n \end{pmatrix}$$

with $b_{ij} = u_iv_j + t_{i,j}$, for $j > i$.

The determinant of B is given by

$$|B| = \frac{u_nv_1}{\prod_{i=1}^{n-1}(-h_{i+1,i})} = (-1)^{n-1}u_nv_1\prod_{i=1}^{n-1} t_{i,i+1}$$

and elements of U and V are

$$u_i = \frac{(-1)^{i-1}}{|H|}|H_{n-i}^{(i)}|\frac{1}{[h_{i+1,i}\cdots h_{n,n-1}]}, \qquad v_j = (-1)^{j-1}|H_{j-1}|[h_{j+1,j}\cdots h_{n,n-1}].$$

For a proof of this Lemma 1 see e.g. [5, 6]. A more detailed proof can be found in [3]. An equivalent lemma can be obtained for lower Hessenberg matrices.

 ISBN: 978-84-616-9216-3

1.2 Unreduced tridiagonal matrices with a finite order

We recall that a tridiagonal matrix having nonzero entries in both the subdiagonal and the superdiagonal is called unreduced tridiagonal matrix. The following result is also well known [5, 6, 3].

Lemma 2 *A nonsingular matrix $H = (h_{ij})_{i,j=1}^n$ is an unreduced tridiagonal matrix if and only if its inverse matrix $B = (b_{ij})_{i,j=1}^n$ has the entries*

$$b_{ij} = \begin{cases} u_i v_j, & for \quad i \geq j; \\ w_i x_j, & for \quad i \leq j. \end{cases}$$

and the entries u_1, v_n, w_n, and x_1 are nonzero entries.

The proof is trivial because a tridiagonal matrix is also lower and upper Hessenberg, and the result follows as an immediate consequence of Lemma 1. Trivially, $u_k v_k = w_k x_k$. If in addition the matrix is symmetric, $u_i = x_i$, and $v_j = w_j$.

Example 1 *We apply Lemma 1 on the following $n \times n$ real symmetric tridiagonal matrix*

$$J_n = \begin{pmatrix} b & a & 0 & 0 & \cdots & 0 \\ a & b & a & 0 & \cdots & 0 \\ 0 & a & b & a & \cdots & 0 \\ 0 & 0 & a & b & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & b \end{pmatrix}.$$

We obtain the matrix

$$B_n = \begin{pmatrix} \frac{|J_{n-1}||J_0|}{|J_n|} & -a\frac{|J_{n-2}||J_0|}{|J_n|} & a^2\frac{|J_{n-3}||J_0|}{|J_n|} & \cdots & (-a)^{n-1}\frac{|J_0||J_0|}{|J_n|} \\ -a\frac{|J_{n-2}||J_0|}{|J_n|} & \frac{|J_{n-2}||J_1|}{|J_n|} & -a\frac{|J_{n-3}||J_1|}{|J_n|} & \cdots & (-a)^{n-2}\frac{|J_1||J_1|}{|J_n|} \\ a^2\frac{|J_{n-3}||J_0|}{|J_n|} & -a\frac{|J_{n-3}||J_1|}{|J_n|} & \frac{|J_{n-3}||J_2|}{|J_n|} & \cdots & (-a)^{n-3}\frac{|J_2||J_2|}{|J_n|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (-a)^{n-1}\frac{|J_0||J_0|}{|J_n|} & (-a)^{n-2}\frac{|J_1||J_1|}{|J_n|} & (-a)^{n-3}\frac{|J_2||J_2|}{|J_n|} & \cdots & \frac{|J_0||J_{n-1}|}{|J_n|} \end{pmatrix}.$$

Because for n determined we expand the determinant by the first column, the involved determinants can easily be computed by using the three-term recurrence relation

$$|J_n| = b|J_{n-1}| - a^2|J_{n-2}|$$

 ISBN: 978-84-616-9216-3

and here we have, in the case $b^2-4a^2\neq 0$*, that the value of determinant is*

$$|J_n|=\frac{\left(b+\sqrt{b^2-4a^2}\right)^{n+1}-\left(b-\sqrt{b^2-4a^2}\right)^{n+1}}{2^{n+1}\sqrt{b^2-4a^2}}$$

that can trivially be obtained using induction, or using the characteristic equation. In the case $b^2-4a^2=0$*, i.e.* $b=\pm 2a$*, we obtain easily* $|J_n|=(n+1)a^n$ *when* $b=2a$ *and* $|J_n|=(n+1)(-a)^n$ *when* $b=-2a$*. Then we have the inverse of the real symmetric matrix* J_n *in all the cases, covering the case when* $a=0$ *which correspond to a matrix* J_n *reduced.*

Some numerical methods for inverting finite Hessenberg matrices and tridiagonal matrices are available, see e.g. [4, 6, 1, 2] and references given there.

2 Inverses of infinite unreduced tridiagonal matrices

We want to extend Lemma 2 to infinite tridiagonal matrices in the next theorem. We recall that if $A=(a_{ij})_{i,j=1}^{\infty}$ is an infinite matrix of complex numbers, the matrix $B=(b_{ij})_{i,j=1}^{\infty}$ is a classical inverse of A if we have $AB=BA=I$.

It is well known that an infinite matrix can have not inverse matrix, for example the matrix corresponding to right-shift operator,

$$S_R=\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

has not associated an inverse matrix, because the product of the first row of S_R to the first column of other matrix can not give 1. It is also well known that an infinite matrix can have two classical inverse matrices, as we will show in some examples, and then infinite many classical inverses, because if B' and B'' are inverses of A, then $\alpha B'+(1-\alpha)B''$ is too an inverse matrix of A, for every $\alpha\in\mathbb{C}$.

Finite tridiagonal matrices are denoted by $\{a_i,b_i,c_i\}$, $1\leq i\leq n$, where the $\{b_i\}_{i=0}^{n}$ are the entries of the principal diagonal. The $\{a_i\}_{i=1}^{n}$ and $\{c_i\}_{i=1}^{n}$ are those of the lower and upper subdiagonal, respectively. We use also the notation $H=\{a_i,b_i,c_i\}$ to indicate the tridiagonal matrix

$$H=\begin{pmatrix} b_0 & c_1 & 0 & 0 & \cdots \\ a_1 & b_1 & c_2 & 0 & \cdots \\ 0 & a_2 & b_2 & c_3 & \cdots \\ 0 & 0 & a_3 & b_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

 ISBN: 978-84-616-9216-3

Theorem 1 *Let H be an infinite invertible matrix. Then H is a tridiagonal unreduced matrix $H = \{a_i, b_i, c_i\}$ if and only if its classical inverse matrix $B = (b_{ij})_{i,j=1}^{\infty}$ has the entries*

$$b_{ij} = \begin{cases} u_i v_j, & \text{for} \quad i \geq j; \\ w_i x_j, & \text{for} \quad i \leq j, \end{cases}$$

where the vectors $U = (u_1, u_2, \ldots), V = (v_1, v_2, \ldots), W = (w_1, w_2, \ldots), X = (x_1, x_2, \ldots)$ satisfy the following recurrence relations

$$\begin{cases} u_2 = \dfrac{1 - b_0 u_1 v_1}{c_1 v_1} \\ u_i = \dfrac{-a_{i-2} u_{i-2} - b_{i-2} u_{i-1}}{c_{i-1}} \end{cases} \qquad \begin{cases} v_2 = \dfrac{-b_0 v_1}{a_1} \\ v_i = \dfrac{-c_{i-2} v_{i-2} - b_{i-2} v_{i-1}}{a_{i-1}} \end{cases}$$

$$\begin{cases} w_2 = \dfrac{-b_0 w_1}{c_1} \\ w_i = \dfrac{-a_{i-2} w_{i-2} - b_{i-2} w_{i-1}}{c_{i-1}} \end{cases} \qquad \begin{cases} x_2 = \dfrac{1 - b_0 w_1 x_1}{a_1 w_1} \\ x_i = \dfrac{-c_{i-2} x_{i-2} - b_{i-2} x_{i-1}}{a_{i-1}} \end{cases}$$

for $i \geq 3$, with $v_1 \neq 0$, $w_1 \neq 0$, and $u_1 v_1 = w_1 x_1$.

Proof Let be H and B the matrices

$$H = \begin{pmatrix} b_0 & c_1 & 0 & 0 & \cdots \\ a_1 & b_1 & c_2 & 0 & \cdots \\ 0 & a_2 & b_2 & c_3 & \cdots \\ 0 & 0 & a_3 & b_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \qquad B = \begin{pmatrix} u_1 v_1 & w_1 x_2 & w_1 x_3 & w_1 x_4 & \cdots \\ u_2 v_1 & u_2 v_2 & w_2 x_3 & w_2 x_4 & \cdots \\ u_3 v_1 & u_3 v_2 & u_3 v_3 & w_3 x_4 & \cdots \\ u_4 v_1 & u_4 v_2 & u_4 v_3 & u_4 v_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $B = UV + T$, $w_{i-1} x_i = u_{i-1} v_i + t_{i-1,i}$, and $t_{i-1,i} = \frac{1}{a_{i,i-1}} \neq 0$.

First we consider the matrix product HB. If we multiply the i-th row of H by the i-th column of B, we have from recurrences,

$$\begin{aligned} a_{i-1} w_{i-1} x_i + b_{i-1} u_i v_1 + c_i u_{i+1} v_i &= a_{i-1}(u_{i-1} v_i + t_{i-1,i}) + b_{i-1} u_i v_i + c_i v_i \frac{-a_{i-1} u_{i-1} - b_{i-1} u_i}{c_i} \\ &= 1 + a_{i-1} u_{i-1} v_i + b_{i-1} u_i v_i - (a_{i-1} u_{i-1} + b_{i-1} u_i) v_i \\ &= 1 + (a_{i-1} u_{i-1} + b_{i-1} u_i - a_{i-1} u_i - b_{i-1} u_i) = 1. \end{aligned}$$

Now, we multiply the i-th row of H by the j-th column of B. In the case $i > j$, we have

$$\begin{aligned} a_{i-1} u_{i-1} v_j + b_{i-1} u_i v_j + c_i u_{i+1} v_j &= \left(a_{i-1} u_{i-1} + b_{i-1} u_i + c_i \frac{-a_{i-1} u_{i-1} - b_{i-1} u_i}{c_i} \right) v_j \\ &= (a_{i-1} u_{i-1} + b_{i-1} u_i - a_{i-1} u_{i-1} - b_{i-1} u_i) v_j = 0. \end{aligned}$$

 ISBN: 978-84-616-9216-3

In the case $i<j$, we have

$$\begin{aligned} a_{i-1}w_{i-1}x_j+b_{i-1}w_ix_j+c_iw_{i+1}x_j &= \left(a_{i-1}w_{i-1}+b_{i-1}w_i+c_i\frac{-a_{i-1}w_{i-1}-b_{i-1}w_i}{c_i}\right)x_j \\ &= (a_{i-1}w_{i-1}+b_{i-1}w_i-a_{i-1}w_{i-1}-b_{i-1}w_i)x_j=0. \end{aligned}$$

We consider now the matrix product BH. If we multiply the i-th row of B by the i-th column of H, we obtain

$$\begin{aligned} u_iv_{i-1}c_{i-1}+u_iv_ib_{i-1}+w_ix_{i+1}a_i &= u_iv_{i-1}c_{i-1}+u_iv_ib_{i-1}+(u_iv_{i+1}+t_{i,i+1})a_i \\ &= u_i(v_{i-1}c_{i-1}+v_ib_{i-1}+v_{i+1}a_i)+t_{i,i+1}a_i \\ &= u_i\left(v_{i-1}c_{i-1}+v_ib_{i-1}+a_i\frac{-c_{i-1}v_{i-1}-b_{i-1}v_i}{a_i}\right)+1=1. \end{aligned}$$

Now, we multiply the i-th row of B by the j-th column of H. In the case $i>j$, we have

$$\begin{aligned} u_iv_{j-1}c_{j-1}+u_iv_jb_{j-1}+u_iv_{j+1}a_j &= u_i\left(v_{j-1}c_{j-1}+v_jb_{j-1}+a_j\frac{-c_{j-1}v_{j-1}-b_{j-1}v_j}{a_j}\right) \\ &= u_i(v_{j-1}c_{j-1}+v_jb_{j-1}-c_{j-1}v_{j-1}-b_{j-1}v_j)=0. \end{aligned}$$

In the case $i<j$, we obtain

$$\begin{aligned} w_ix_{j-1}c_{j-1}+w_ix_jb_j+w_ix_{j+1}a_j &= w_i\left(x_{j-1}c_{j-1}+x_jb_{j-1}+a_j\frac{-c_{j-1}x_{j-1}-b_{j-1}x_{j-1}}{a_j}\right) \\ &= w_i(x_{j-1}c_{j-1}+x_jb_{j-1}-c_{j-1}x_{j-1}-b_{j-1}x_{j-1})=0. \end{aligned}$$

Therefore, matrix B is the classical inverse of the matrix H and conversely, matrix H is the classical inverse of the matrix B.

We must to prove that the condition $u_1v_1=w_1x_1$ implies $u_kv_k=w_kx_k$, for $k=2,3,\ldots$ Indeed, for $k=2$,

$$\begin{aligned} u_2v_2 &= \frac{1-b_0u_1v_1}{c_1v_1}\cdot\frac{-b_0v_1}{a_1}=\frac{-b_0(1-b_0u_1v_1)}{a_1c_1}=\frac{-b_0(1-b_0w_1x_1)}{a_1c_1} \\ &= \frac{1-b_0w_1x_1}{a_1w_1}\cdot\frac{-b_0w_1}{c_1}=x_2w_2=w_2x_2. \end{aligned}$$

In addition,

$$\begin{aligned} a_1b_1u_1v_2+b_1c_1u_2v_1 &= a_1b_1u_1\frac{-b_0v_1}{a_1}+b_1c_1v_1\frac{1-b_0u_1v_1}{c_1v_1} \\ &= -b_0b_1u_1v_1+b_1(1-b_0u_1v_1)=-b_0b_1w_1x_1+b_1(1-b_0w_1x_1) \\ &= b_1c_1\frac{-b_0w_1}{c_1}x_1+a_1b_1w_1\frac{1-b_0w_1x_1}{a_1w_1} \\ &= b_1c_1w_2x_1+a_1b_1w_1x_2=a_1b_1w_1x_2+b_1c_1w_2x_1. \end{aligned}$$

 ISBN: 978-84-616-9216-3

Now we use induction,

$$u_{k-1}v_{k-1} = w_{k-1}x_{k-1}, \qquad u_kv_k = w_kx_k$$

In addition,

$$a_{k-1}b_{k-1}u_{k-1}v_k + b_{k-1}c_{k-1}u_kv_{k-1} = a_{k-1}b_{k-1}w_{k-1}x_k + b_{k-1}c_{k-1}w_kx_{k-1}.$$

Therefore,

$$\begin{aligned} u_{k+1}v_{k+1} &= \frac{-a_{k-1}u_{k-1} - b_{k-1}u_k}{c_k} \cdot \frac{-c_{k-1}v_{k-1} - b_{k-1}v_k}{a_k} \\ &= \frac{1}{a_kc_k}(a_{k-1}c_{k-1}w_{k-1}x_{k-1} + a_{k-1}b_{k-1}w_{k-1}x_k + b_{k-1}c_{k-1}w_kx_{k-1} + b_{k-1}b_{k-1}w_kx_k) \\ &= \frac{-a_{k-1}w_{k-1} - b_{k-1}w_k}{c_k} \cdot \frac{-c_{k-1}x_{k-1} - b_{k-1}x_k}{a_k} = w_{k+1}x_{k+1}. \end{aligned}$$

The proof is complete, $u_{k+1}v_{k+1} = w_{k+1}x_{k+1}, \forall k$. □

Remark 1 *In the establishment of the preceding theorem, we have recurrence relations and the conditions $v_1 \neq 0$, $w_1 \neq 0$, and $u_1v_1 = w_1x_1$. Then, three of these parameters are free. When we choose different values of these, we obtain different inverse matrices of the infinite tridiagonal matrix H.*

Example 2 *We illustrate with the infinite real symmetric tridiagonal matrix*

$$H = \begin{pmatrix} 1 & \frac{-2}{5} & 0 & \cdots \\ \frac{-2}{5} & 1 & \frac{-2}{5} & \cdots \\ 0 & \frac{-2}{5} & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $b_i = 1$ and $a_i = c_i = \frac{-2}{5}$. If we choose, say $u_1 = v_1 = 1$, we have

$$U = \left(1, 0, -1, \frac{-5}{2}, \frac{-21}{4}, \frac{-85}{8}, \cdots\right)^t, \qquad V = \left(1, \frac{5}{2}, \frac{21}{4}, \frac{85}{8}, \frac{341}{16}, \frac{1365}{32}, \cdots\right).$$

We obtain by symmetry B' as a classical inverse of matrix H. However, if we choose, say $u_1 = \frac{5}{4}$ and $v_1 = 1$, we have

$$U = \left(\frac{5}{4}, \frac{5}{8}, \frac{5}{16}, \frac{5}{32}, \frac{5}{64}, \frac{5}{128}, \cdots\right)^t, \qquad V = \left(1, \frac{5}{2}, \frac{21}{4}, \frac{85}{8}, \frac{341}{16}, \frac{1365}{32}, \cdots\right).$$

We obtain B'' as an inverse of H. Finally, if we choose, say $u_1 = 0$ and $v_1 = 1$, we have

$$U = \left(0, \frac{-5}{2}, \frac{-25}{4}, \frac{-105}{8}, \frac{-425}{16}, \frac{-1705}{32}, \cdots\right)^t, \qquad V = \left(1, \frac{5}{2}, \frac{21}{4}, \frac{85}{8}, \frac{341}{16}, \frac{1365}{32}, \cdots\right),$$

 ISBN: 978-84-616-9216-3

and we obtain B''' as an inverse of H. The matrices B', B'', and B''' are

$$B'=\begin{pmatrix} 1 & 0 & -1 & \frac{-5}{2} & \frac{-21}{4} & \cdots \\ 0 & 0 & \frac{-5}{2} & \frac{-25}{4} & \frac{-105}{8} & \cdots \\ -1 & \frac{-5}{2} & \frac{-21}{4} & \frac{-105}{8} & \frac{-441}{16} & \cdots \\ \frac{-5}{2} & \frac{-25}{4} & \frac{-105}{8} & \frac{-425}{16} & \frac{-1785}{32} & \cdots \\ \frac{-21}{4} & \frac{-105}{8} & \frac{-441}{16} & \frac{-1785}{32} & \frac{-7161}{64} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, B''=\begin{pmatrix} \frac{5}{4} & \frac{5}{8} & \frac{5}{16} & \frac{5}{32} & \frac{5}{64} & \cdots \\ \frac{5}{8} & \frac{25}{16} & \frac{25}{32} & \frac{25}{64} & \frac{25}{128} & \cdots \\ \frac{5}{16} & \frac{25}{32} & \frac{105}{64} & \frac{105}{128} & \frac{105}{256} & \cdots \\ \frac{5}{32} & \frac{25}{64} & \frac{105}{128} & \frac{425}{256} & \frac{425}{512} & \cdots \\ \frac{5}{64} & \frac{25}{128} & \frac{105}{256} & \frac{425}{512} & \frac{1705}{1024} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

$$B'''=\begin{pmatrix} 0 & \frac{-5}{2} & \frac{-25}{4} & \frac{-105}{8} & \frac{-425}{16} & \cdots \\ \frac{-5}{2} & \frac{-25}{4} & \frac{-125}{8} & \frac{-525}{16} & \frac{-2125}{32} & \cdots \\ \frac{-25}{4} & \frac{-125}{8} & \frac{-525}{16} & \frac{-2205}{32} & \frac{-8925}{64} & \cdots \\ \frac{-105}{8} & \frac{-525}{16} & \frac{-2205}{32} & \frac{-8925}{64} & \frac{-36125}{128} & \cdots \\ \frac{-425}{16} & \frac{-2125}{32} & \frac{-8925}{64} & \frac{-36125}{128} & \frac{-144925}{256} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

These three matrices are examples of classical inverses of matrix H. Note also that if we choose $v_1 = i$, we obtain the vector $V = \left(i, \frac{5i}{2}, \frac{21i}{4}, \frac{85i}{8}, \frac{341i}{16}, \ldots\right)$, and taking $u_1 = 1$, for example, we have a complex classical inverse of a real matrix H.

Example 3 *We consider the infinite, real or complex, symmetric tridiagonal matrix*

$$H = \begin{pmatrix} 1+a^2 & a & 0 & \cdots \\ a & 1+a^2 & a & \cdots \\ 0 & a & 1+a^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

then $b_i = 1 + a^2$ and $a_i = c_i = a$. For $u_1 = v_1 = 1$, we have the following vectors $U = (1, -a, a^2, -a^3, a^4, ...)^t$ and

$$V = \left(\frac{1-a^2}{1-a^2}, -\frac{1-a^4}{a(1-a^2)}, \frac{1-a^6}{a^2(1-a^2)}, -\frac{1-a^8}{a^3(1-a^2)}, \frac{1-a^{10}}{a^4(1-a^2)}, \ldots\right).$$

We obtain by symmetry the matrix

$$B = \frac{1}{1-a^2}\begin{pmatrix} 1-a^2 & -a(1-a^2) & a^2(1-a^2) & -a^3(1-a^2) & a^4(1-a^2) & \cdots \\ -a(1-a^2) & 1-a^4 & -a(1-a^4) & a^2(1-a^4) & -a^3(1-a^4) & \cdots \\ a^2(1-a^2) & -a(1-a^4) & 1-a^6 & -a(1-a^6) & a^2(1-a^6) & \cdots \\ -a^3(1-a^2) & a^2(1-a^4) & -a(1-a^6) & 1-a^8 & -a(1-a^8) & \cdots \\ a^4(1-a^2) & -a^3(1-a^4) & a^2(1-a^6) & -a(1-a^8) & 1-a^{10} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

 ISBN: 978-84-616-9216-3

a classical inverse matrix of the matrix H. Note that if we take $a = i$, we have a classical inverse of the bidiagonal complex matrix $H = \{i, 0, i\}$. If we take $a = 0$, we have I as the inverse of the matrix I. Finally, taking $a = 1$ and after simplifying, the inverse of the matrix $H = \{1, 2, 1\}$ is given by

$$B = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & \cdots \\ -1 & 2 & -2 & 2 & -2 & \cdots \\ 1 & -2 & 3 & -3 & 3 & \cdots \\ -1 & 2 & -3 & 4 & -4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} = ((-1)^{i+j} \min\{i, j\}_{i,j=1}^{\infty}).$$

Theorem 1 gives us classical inverses of a tridiagonal matrix H. Conversely, we show in the next corollary the way for obtaining the inverse of an infinite matrix B with structure UV and WX.

Corollary 1 *Let $B = (b_{ij})_{i,j=1}^{\infty}$ be an infinite invertible matrix with structure UV for $i \geq j$, and WX for $i \leq j$. Its classical inverse, the infinite tridiagonal matrix $H = \{a_i, b_i, c_i\}$, is unique, The entries of the inverse matrix H are given by the following recursive relations:*

$$b_0 = \frac{b_{22}}{b_{11}b_{22} - b_{12}b_{21}}, \quad b_{i-2} = \frac{b_{i-2,1}(b_{i-1,i-1}b_{1,i}b_{i,1} - b_{ii}b_{1,i-1}b_{i-1,1}}{c_{i-2}b_{i-1,1}(b_{ii}b_{1,i-2}b_{i-2,1} - b_{i-2,i-2}b_{1,i}b_{i,1})}, \text{ and}$$

$$\begin{cases} a_1 = \dfrac{1 - b_0x_1}{x_2} \\ a_{i-1} = \dfrac{-c_{i-2}x_{i-2} - b_{i-2}x_{i-1}}{x_i} \end{cases}, \text{ and } \begin{cases} c_1 = \dfrac{1 - b_0u_1}{u_2} \\ c_{i-1} = \dfrac{-a_{i-2}u_{i-2} - b_{i-2}u_{i-1}}{u_i} \end{cases},$$

for $i \geq 3$. The order in the computation of parameters is the following: first b_0, after a_1 and c_1, after b_1, after a_2 and c_2 and sequentially.

Proof We take $v_1 = w_1 = 1$ without loss of generality. The matrix B is

$$B = \begin{pmatrix} u_1 & x_2 & x_3 & x_4 & \cdots \\ u_2 & u_2v_2 & w_2x_3 & w_2x_4 & \cdots \\ u_3 & u_3v_2 & u_3v_3 & w_3x_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

with $x_1 = u_1$ and $u_kv_k = w_kx_k$, $k = 2, 3, \ldots$ The first column of B is the vector U and the first row of B is the vector X. Thus u_i, x_i, $i = 1, 2, \ldots$ are known. Recurrence relations of a_i and c_i are obtained from recurrences of x_i and u_i from Theorem 1. We must determine b_0 and b_i. Entry b_{22} of the matrix B is u_2v_2. Since u_2 is known, we have $v_2 = \frac{b_{22}}{u_2} = \frac{b_{22}}{b_{21}}$. Recurrence for v_i, with $v_1 = 1$, allows us to obtain

$$b_0 = -a_1v_2 = \frac{b_0x_1 - 1}{x_2}v_2 \quad \Rightarrow \quad b_0x_2 = b_0x_1v_2 - v_2 \quad \Rightarrow \quad b_0(x_1v_2 - x_2) = v_2 \quad \Rightarrow$$

 ISBN: 978-84-616-9216-3

$$b_0 = \frac{v_2}{x_1 v_2 - x_2} = \frac{\frac{b_{22}}{b_{21}}}{b_{11}\frac{b_{22}}{b_{21}} - b_{12}} = \frac{b_{22}}{b_{11}b_{22} - b_{12}b_{21}},$$

where the denominator is nonzero because, in other case, the two first columns and rows of the matrix B are proportional, and matrix B is not invertible.

We give some details about the computations of the b_i. Since

$$a_{i-1} = \frac{-c_{i-2}v_{i-2} - b_{i-2}v_{i-1}}{v_i} = \frac{-c_{i-2}x_{i-2} - b_{i-2}x_{i-1}}{x_i}, \text{ we obtain}$$

$$-c_{i-2}v_{i-2}x_i - b_{i-2}v_{i-1}x_i = -c_{i-2}x_{i-2}v_i - b_{i-2}x_{i-1}v_i \quad \Rightarrow \quad b_{i-2} = \frac{v_{i-1}x_i - x_{i-1}v_i}{c_{i-2}(x_{i-2}v_i - v_{i-2}x_i)}.$$

Furthermore, since $u_i v_i = b_{ii} \quad \Rightarrow \quad v_i = \frac{b_{ii}}{u_i} = \frac{b_{ii}}{b_{i1}}$, we have

$$b_{i-2} = \frac{\frac{b_{i-1,i-1}}{b_{i-1,1}}b_{1i} - b_{1,i-1}\frac{b_{ii}}{b_{i1}}}{c_{i-2}\left(b_{1,i-2}\frac{b_{ii}}{b_{i1}} - \frac{b_{i-2,i-2}}{b_{i-2,1}}b_{1i}\right)} = \frac{b_{21}(b_{i-1,i-1}b_{1i}b_{i1} - b_{ii}b_{1,i-1}b_{i-1,1})}{c_{i-2}b_{i-1,1}(b_{ii}b_{1,i-2}b_{i-2,1} - b_{i-2,i-2}b_{1i}b_{i1})},$$

where numerator and denominator are nonzero because matrix B is invertible. In particular, for $i = 3$ we obtain

$$b_1 = \frac{b_{11}(b_{22}b_{13}b_{31} - b_{33}b_{12}b_{21})}{c_1 b_{21}(b_{33}b_{11}b_{11} - b_{11}b_{13}b_{31})} = \frac{b_{22}b_{13}b_{31} - b_{33}b_{12}b_{21}}{c_1 b_{21}\,(b_{11}b_{33} - b_{13}b_{31})}.$$

Products of rows by columns of H and B are already computed in the proof of Theorem 1. In summary, H is the inverse matrix of B, and conversely. The unicity of matrix H follows from the unicity of the expressions for a_i, b_i, and c_i. □

3 Inverses of infinite reduced tridiagonal matrices

When a tridiagonal matrix H has at least a null entry on its subdiagonal, we can calculate its classical inverse in a similar way as the known method of the Schur complement for a matrix of finite order.

Proposition 1 *Let H be an infinite invertible tridiagonal matrix with only a zero entry on its subdiagonal and nonzero entries on its superdiagonal. Then its classical inverse matrix can be calculate using a block matrix procedure. If matrix $H = \left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline 0 & H_{22} \end{array}\right)$, a classical inverse matrix is $B = \left(\begin{array}{c|c} H_{11}^{-1} & -H_{11}^{-1}H_{12}H_{22}^{-1} \\ \hline 0 & H_{22}^{-1} \end{array}\right)$, where H_{11} is a finite nonsingular unreduced tridiagonal matrix and H_{22} is an infinite invertible unreduced tridiagonal matrix.*

Proof The proof follows trivially by considering the products $HB = I$ and $BH = I$. Thus matrix B is a classical inverse of matrix H. □

 ISBN: 978-84-616-9216-3

Example 4 *Let H be the infinite tridiagonal matrix, with an unique null entry on its subdiagonal, $h_{43} = 0$, a classical inverse of H will be B, where*

$$H = \left(\begin{array}{ccc|ccc} 2 & 1 & 0 & & & \cdots \\ 1 & 2 & 1 & & & \cdots \\ 0 & 1 & 2 & 1 & & \cdots \\ \hline & & 0 & 2 & 1 & \cdots \\ & & & 1 & 2 & 1 & \cdots \\ & & & & 1 & 2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right), \quad B = \left(\begin{array}{ccc|ccc} \frac{3}{4} & \frac{-1}{2} & \frac{1}{4} & & B_{12} & & \cdots \\ \frac{-1}{2} & 1 & \frac{-1}{2} & & & & \cdots \\ \frac{1}{4} & \frac{-1}{2} & \frac{3}{4} & & & & \cdots \\ \hline 0 & 0 & 0 & 1 & -1 & 1 & \cdots \\ 0 & 0 & 0 & -1 & 2 & -2 & \cdots \\ 0 & 0 & 0 & 1 & -2 & 3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right).$$

The block matrix entry B_{12} is

$$B_{12} = -H_{11}^{-1}H_{12}H_{22}^{-1} = -\begin{pmatrix} \frac{3}{4} & \frac{-1}{2} & \frac{1}{4} \\ \frac{-1}{2} & 1 & \frac{-1}{2} \\ \frac{1}{4} & \frac{-1}{2} & \frac{3}{4} \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \end{pmatrix}\begin{pmatrix} 1 & -1 & 1 & \cdots \\ -1 & 2 & -2 & \cdots \\ 1 & -2 & 3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$= -\begin{pmatrix} \frac{1}{4} & 0 & 0 & \cdots \\ \frac{-1}{2} & 0 & 0 & \cdots \\ \frac{3}{4} & 0 & 0 & \cdots \end{pmatrix}\begin{pmatrix} 1 & -1 & 1 & \cdots \\ -1 & 2 & -2 & \cdots \\ 1 & -2 & 3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \frac{-1}{4} & \frac{1}{4} & \frac{-1}{4} & \cdots \\ \frac{1}{2} & \frac{-1}{2} & \frac{1}{2} & \cdots \\ \frac{-3}{4} & \frac{3}{4} & \frac{-3}{4} & \cdots \end{pmatrix}.$$

The case of finitely many zeros on the subdiagonal follows from Proposition 1.

Corollary 2 *Let H be an infinite invertible tridiagonal matrix with a finite number of zero entries on its subdiagonal and nonzero entries on its superdiagonal. Then its classical inverse matrices can be calculate using the same block matrix procedure as given in Proposition 1, but now the finite nonsingular tridiagonal matrix H_{11} is reduced.*

This corollary can be extended in a natural way when the tridiagonal matrix H has infinitely many zeros on its subdiagonal. An analogous procedure is valid if the null entries are on its superdiagonal.

4 Conclusions

We have proposed a method for building classical inverses of a general, unreduced as well as reduced, infinite (real or complex) tridiagonal matrix. Some free parameters have been chosen and different inverses have been obtained using the recurrence relations involved. Conversely, a constructive method was presented for calculating the inverse of an infinite matrix with the structure UV and WX.

 ISBN: 978-84-616-9216-3

References

[1] J. Abderramán Marrero, M. Rachidi, V. Tomeo, *Non-symbolic algorithms for the inversion of tridiagonal matrices*, J. Comp. Appl. Math. **252** (2013) 3–11.

[2] J. Abderramán Marrero, M. Rachidi, V. Tomeo, *On new algorithms for inverting Hessenberg matrices*, J. Comp. Appl. Math. **252** (2013) 12–20.

[3] J. Abderramán Marrero, V. Tomeo, E. Torrano, *A new tool for generating orthogonal polynomial sequences*, Proceedings of the 13th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE **1** (2013) 25–36.

[4] B. Bukhberger, G. A. Emel'yanenko, *Methods of inverting tridiagonal matrices*, Comput. Math. Math. Phys. URSS **13** (1973) 10–20.

[5] D. K. Faddeev, *Properties of a matrix, inverse of a Hessenbeerg matrix*, Journal Mathematical Sciences **24** (1984) 118–120.

[6] Y. Ikebe, *On inverses of Hessenberg matrices*, Linear Algebra Appl. **24** (1979) 93–97.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

On the construction of finite-term recursions for Fractional Differential Equations

Lidia Aceto[1], Cecilia Magherini[1] and Paolo Novati[2]

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

[2] *Department of Mathematics, University of Padova*

emails: aceto@dm.unipi.it, magherini@dm.unipi.it, novati@math.unipd.it

Abstract

This paper deals with the numerical solution of Fractional Differential Equations by means of m-step recursions. For the construction of such formulas, we consider a technique based on a rational approximation of the generating functions of Fractional Backward Differentiation Formulas (FBDFs). The so-defined methods simulate very well the properties of the underlying FBDFs with noticeable advantages in terms of memory saving. This fact becomes particularly evident when they are used for discretizing fractional partial differential equations like the ones occurring in some population dynamic models.

Key words: Fractional Differential Equations, Fractional BDFs, Matrix functions

1 Introduction

We consider the numerical solutions of Fractional Differential Equations (FDEs) of the type

$$ {}_{t_0}D_t^{\alpha} y(t) = g(t, y(t)), \qquad t \in (t_0, T], \quad 0 < \alpha < 1, \tag{1} $$

where ${}_{t_0}D_t^{\alpha}$ denotes the fractional derivative operator in the Caputo sense (see, e.g., [2]) defined as

$$ {}_{t_0}D_t^{\alpha} y(t) = \frac{1}{\Gamma(1-\alpha)} \int_{t_0}^{t} \frac{y'(u)}{(t-u)^{\alpha}} du, \tag{2} $$

being $\Gamma(\cdot)$ the standard Gamma function. As well known, the use of the Caputo's definition for the fractional derivative allows to treat the initial conditions at t_0 for FDEs in the

 ISBN: 978-84-616-9216-3

same manner as for integer order differential equations. Setting $y(t_0) = y_0$ the solution of (1) exists and is unique under the hypothesis that g is continuous and fulfils a Lipschitz condition with respect to the second variable.

A classical approach for solving (1) is based on the discretization of the fractional derivative (2) which leads to the so-called Fractional Backward Differentiation Formulas (FBDFs) introduced in [1]. In more details, FBDFs are given by the full-term recursions

$$\sum\nolimits_{j=0}^{n} \omega_{n-j} y_j = h^\alpha g(t_n, y_n), \quad p \le n \le N, \tag{3}$$

where $h = (T - t_0)/N$ is the step-length of the uniform meshgrid $t_j = t_0 + jh$, $y_j \approx y(t_j)$ and ω_{n-j} are the Taylor coefficients of the generating function

$$\omega_p^{(\alpha)}(\zeta) = \left(a_0 + a_1\zeta + ... + a_p\zeta^p\right)^\alpha, \quad 1 \le p \le 6; \tag{4}$$

here $\{a_0, a_1, \ldots, a_p\}$ are the coefficients of the underlying BDF.

2 The fractional derivative approximation

Our idea is to design new schemes based on a rational approximation of (4)

$$R_m(\zeta) \approx \omega_p^{(\alpha)}(\zeta), \quad R_m(\zeta) = \frac{p_m(\zeta)}{q_m(\zeta)}, \quad p_m, q_m \in \Pi_m, \tag{5}$$

where Π_m denotes the set of polynomials of degree not exceeding m. Writing $p_m(\zeta) = \sum_{i=0}^{m} \alpha_i \zeta^i$ and $q_m(\zeta) = \sum_{i=0}^{m} \beta_i \zeta^i$, the above approximation naturally leads to implicit m-step recursions of the type

$$\sum\nolimits_{j=n-m}^{n} \alpha_{n-j} y_j = h^\alpha \sum\nolimits_{j=n-m}^{n} \beta_{n-j} g(t_j, y_j), \quad n \ge m. \tag{6}$$

After considering a BDF discretization of order p of the first derivative operator, which can be represented by a lower triangular banded Toeplitz matrix of the type

$$A_p = \begin{pmatrix} a_0 & 0 & & & 0 \\ \vdots & a_0 & 0 & & \\ a_p & & \ddots & 0 & \\ 0 & \ddots & & \ddots & 0 \\ & 0 & a_p & \cdots & a_0 \end{pmatrix} \in \mathbb{R}^{N \times N}, \tag{7}$$

we approximate the Caputo's fractional differential operator ${}_{t_0}D_t^\alpha$ by calculating A_p^α. The technique we have used is based on the fact that the first column of A_p^α contains the first

 ISBN: 978-84-616-9216-3

N coefficients of the Taylor expansion of $\omega_p^{(\alpha)}(\zeta)$. In particular, a k-point Gauss-Jacobi quadrature rule has been used for approximating the contour integral representation of the function z^α. The obtained schemes are of type (6) with $m = kp \ll N$ and they generalize in some sense the methods based on the Short Memory Principle in which the truncated Taylor expansion of (4) is considered (see [2, §8.3] for some examples). The advantages in terms of memory saving are noticeable especially in the case when (1) arises from the semi-discretization of fractional partial differential equations.

3 A numerical example

We consider the following problem

$$\begin{aligned} {}_0D_t^\alpha u(x,t) &= \frac{\partial(p(x)u(x,t))}{\partial x} + K_\alpha \frac{\partial^2 u(x,t)}{\partial x^2} + ru(x,t)\left(1 - \frac{u(x,t)}{K}\right), \\ u(0,t) &= u(5,t) = 0, \quad t \in [0,1], \\ u(x,0) &= x^2(5-x)^2, \quad x \in [0,5]. \end{aligned}$$

This is a particular instance of the time-fractional Fokker-Planck equation with a nonlinear source term [3]. In population dynamics, its solution $u(x,t)$ represents the population density at location x and time t and the nonlinear source term in the equation is known as Fisher's growth term. The application of the classical second order semi-discretization in space leads to a nonhomogeneous nonlinear fractional differential problem of type (1) which has been solved by applying the schemes proposed. The results of our experiments are very encouraging.

Acknowledgements

This work has been supported by the INdAM-GNCS Project 2014 "Metodi numerici per modelli di propagazione di onde elettromagnatiche in tessuti biologici".

References

[1] C. Lubich, *Discretized fractional calculus*, SlAM J. Math. Anal. **17** (1986) 704–719.

[2] I. Podlubny, *Fractional differential equations*, Mathematics in Science and Engineering, 198. Academic Press, Inc., San Diego, CA, 1999.

[3] Q. Yang, F. Liu, I. Turner, *Stability and convergence of an effective numerical method for the Time-Space Fractional Fokker-Planck Equation with a nonlinear source term*, Int. J. Differ. Equ., Art. ID 464321 (2010) 22 pp..

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

If "football fever could be a dose of dengue", the "Simon Hay fever" should have given a dose of samba

Maira Aguiar[1], Filipe Rocha[1] and Nico Stollenwerk[1]

[1] *Centro de Matemática e Aplicações Fundamentais, Universidade de Lisboa, Portugal*

emails: mairа@ptmat.fc.ul.pt, frocha@ptmat.fc.ul.pt, nico@ptmat.fc.ul.pt

Abstract

An opinion published in Nature [1] has stated that dengue fever could be a significant problem in some of the Brazilian cities hosting the games, however, the conclusions were taken after a brief observation of the available data, analyzing its mean and standard deviation only, without a major scientific reason to cause alarm for the world cup in Brazil. Here, with a more careful data analysis for the Brazilians cities hosting the games, we show that the seasonality of the disease plays a major role in transmission. The density of dengue cases is residual during the winter period and the fans of football are not likely to get dengue during the tournament season.

Key words: Dengue fever, Data analysis, time series, box-plot

1 Introduction

Dengue is a viral mosquito-borne infection, a leading cause of illness and death in the tropics and subtropics. It is estimated that every year 390 million dengue infections per year, of which 96 million manifest apparently (any level of disease severity). [2]. In many countries in Asia and South America dengue fever/dengue hemorrhagic fever (DF/DHF) has become a substantial public health concern leading to serious social-economic costs. The infection can be asymptomatic or show with a broad clinical spectrum. There is no specific treatment for dengue, and a vaccine which simulates a protective immune response to all four serotypes is not yet available. Tetravalent vaccines are under investigation, but so far, prevention of exposure remains the only alternative to prevent dengue transmission.

Dengue fever epidemiology dynamics shows large fluctuations of disease incidence and mathematical models describing transmission of disease ultimately aim to be used as predictive tools to evaluate the introduction of intervention strategies [3, 4].

In Brazil, the occurrence of the dengue fever is persisting and is increasing since 1980s. By 2000, dengue virus (DENV) transmission was reported in 22 of 27 Brazilian states, occupying a significant place in the international ranking for total cases of the disease, according to the World Health Organization (WHO) [5, 6]. The disease outbreak starts during the rainy season, from Mid of September till Mid of May (see Fig.3) where vector infestation increase considerably. The suspected dengue cases are of compulsory notification and all reported cases from public health services or private health providers are included in the notification database (SINAN), which is openly accessible via the internet [7].

This years World Cup 2014 tournament will be held in Brazil, during the winter season, starting on June 12 and ending on July 13. According to estimates from the Brazilian Tourism Ministry, more than 600,000 football fans will visit the country. The World Cup will be staged across twelve host cities in Brazil: Belo Horizonte, Braslia, Cuiab, Curitiba, Fortaleza, Manaus, Natal, Porto Alegre, Recife, Rio de Janeiro, Salvador and So Paulo, and according to the opinion by Simon Hay, published in Nature [1], dengue fever could be a significant problem in Fortaleza, Natal and Salvador. The authors claimed that much could be done by the authorities there to reduce dengue risk in the run-up to the tournament and have advised travelers to "select accommodation with screened windows and doors and air conditioning; use insecticides indoors; wear clothing that covers the arms and legs, especially during early morning and late afternoon, when the chance of being bitten is greatest; and apply insect repellent to clothing and exposed skin".

These conclusions were taken after the authors have analyzed the mean and the standard deviation for the available data of the twelve cities in Brazil that will host the games. There was not a major scientific reason to cause fear or alarm for the world cup, and specially also to promote expensive accommodations where screened windows and doors and air conditioning is required, for example. This opinion by Simon Hay had a big repercussion in the media worldwide, but will dengue be effectively a threat during the football tournament?

A systematic data collection and a correct analysis should be performed in order to minimizing the false predictions that could be generated by using wrong data or its misinterpretation [4]. In this manuscript we perform a more careful data analysis for the Brazilians cities that are hosting the games and we show that the risk of being infected by a dengue virus is seasonal and also proportional to the population density in Brazil, increasing during the rainy season and the presence of vector and human population density.

2 Methods and Results

In this study, we analyzed the available epidemiological dengue data for the Brazilians cities which are going to host the football games during the FIFA World Cup in 2014. A time series and a box-plot analysis were performed.

The epidemiological dengue data was obtained in the Brazilian notification database

 ISBN: 978-84-616-9216-3

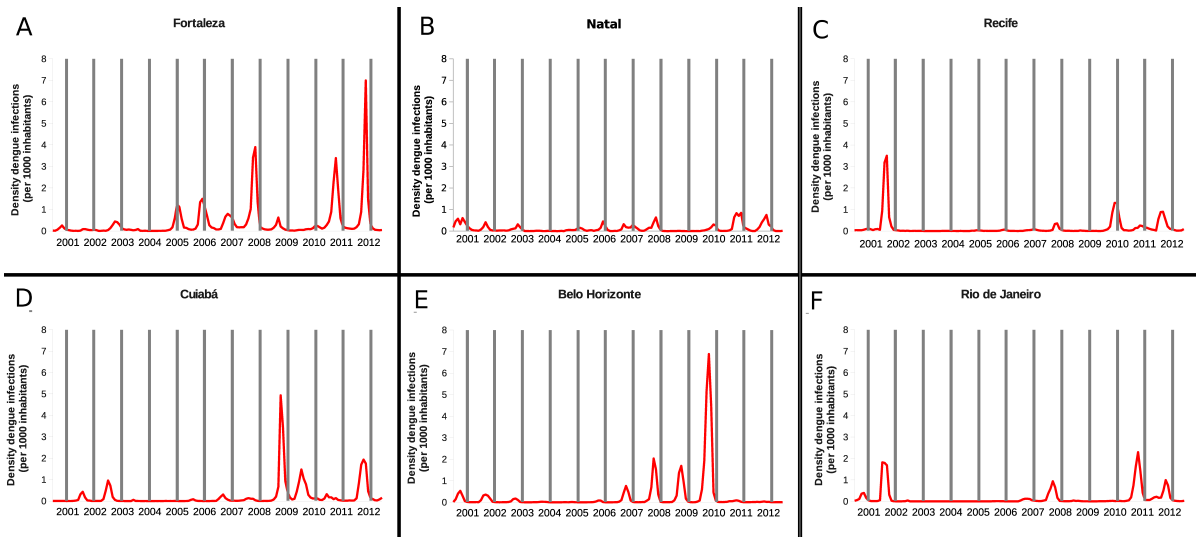


Figure 1: Time series of number of dengue cases per 1000 inhabitants, from 2001 to 2012, for six of the Brazilian cities hosting the football games during the World Cup 2014. In A) Fortaleza, in B) Natal, in C) Recife, in D) Cuiabá, in E) Belo Horizonte, and in F) Rio de Janeiro. The white bars confine the the period where the World Cup is scheduled to happen, from 12 of June to 13 of July.

SINAN [7], the same data set used in [1]. The monthly number dengue cases that were confirmed in laboratory are available from 2001 to 2012 only, and since the population density can be different for each one of the Brazilian cities, we assume that the disease transmission is density-dependent.

The information given by the precipitation data in Brazil was obtained from the National Institute of Meteorology (INMET)[8], where a meteorological data base for research is available. The data base holds a daily weather data in digital form, such as historical series of INMET network stations. Climatological data, combined with the epidemiological data analysis, were used to conclude that the fans of football are not likely to get dengue during the tournament season.

2.1 Time Series Analysis

The epidemiological data were analyzed taking into consideration the human population density for each one of the cities, and the conclusions were taken based on the results obtained by a time series and box plot analysis, combined with the information given by the precipitation data.

Time series of density dengue infection are shown in Fig.1, for some of the Brazilian

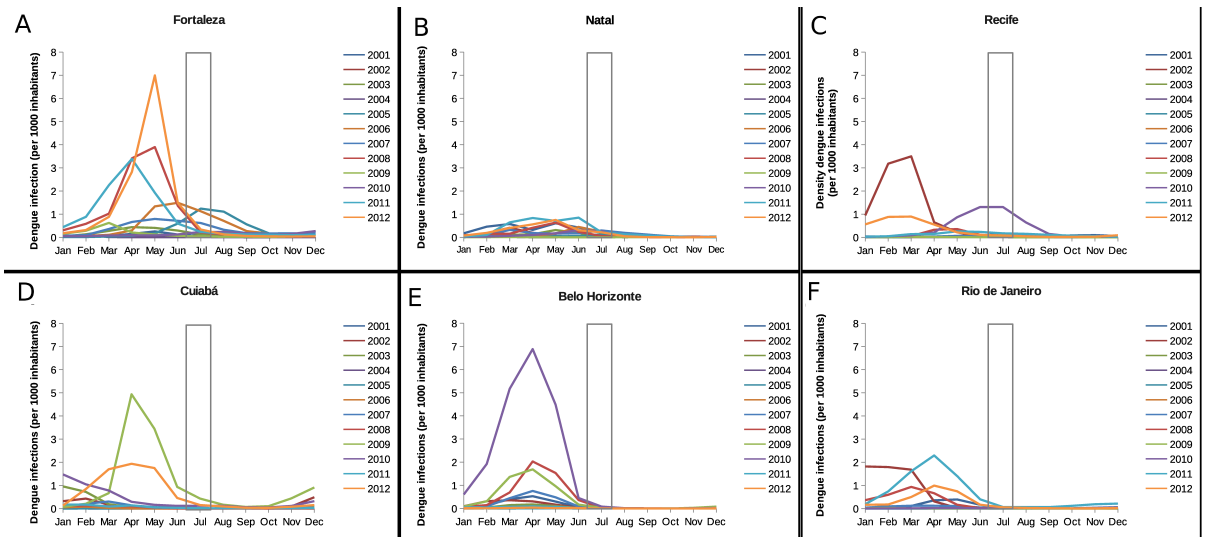


Figure 2: Time series of number of dengue cases per 1000 inhabitants, from 2001 to 2012, for six of the Brazilian cities hosting the football games during the World Cup 2014. Here we plot the number of cases per month, showing the dynamics of the dengue epidemics in each year. In A) Fortaleza, in B) Natal, in C) Recife, in D) Cuiabá, in E) Belo Horizonte, and in F) Rio de Janeiro. The white bars confine the the period where the World Cup is scheduled to happen, from 12 of June to 13 of July.

cities which are hosting the World Cup football games. Among the twelve selected cities, Fortaleza (1A), Cuiabá (1D) and Belo Horizonte (1E) appears to be the cities with higher density of dengue cases. Manaus, Recife (1C), Brasília and Rio de Janeiro (1F) have shown a mild density of cases with rarely high outbreaks during the last 12 years. The density of cases is very small and not significant in Natal (1B), Salvador and São Paulo, and for Curitiba and Porto Alegre, the density of cases are negligible, with only few occasional notified and confirmed cases.

In Fig.2 we plot twelve years of monthly data, for each one of the cities we are studying. We observe that, for all the twelve cities, the dengue season starts in January/February, with the peak of the epidemics around March/April. In May, the number of cases have decreased considerably, and in June and July, during the football games period (signalized by gray bars in the graphics), the number of cases are residual. This pattern is also confirmed to happen for the other cities that are also hosting the football games [10] and it is coherent with the increase of vector infestation, that is highly correlated with the rainfall [9] (see Fig.3).

 ISBN: 978-84-616-9216-3

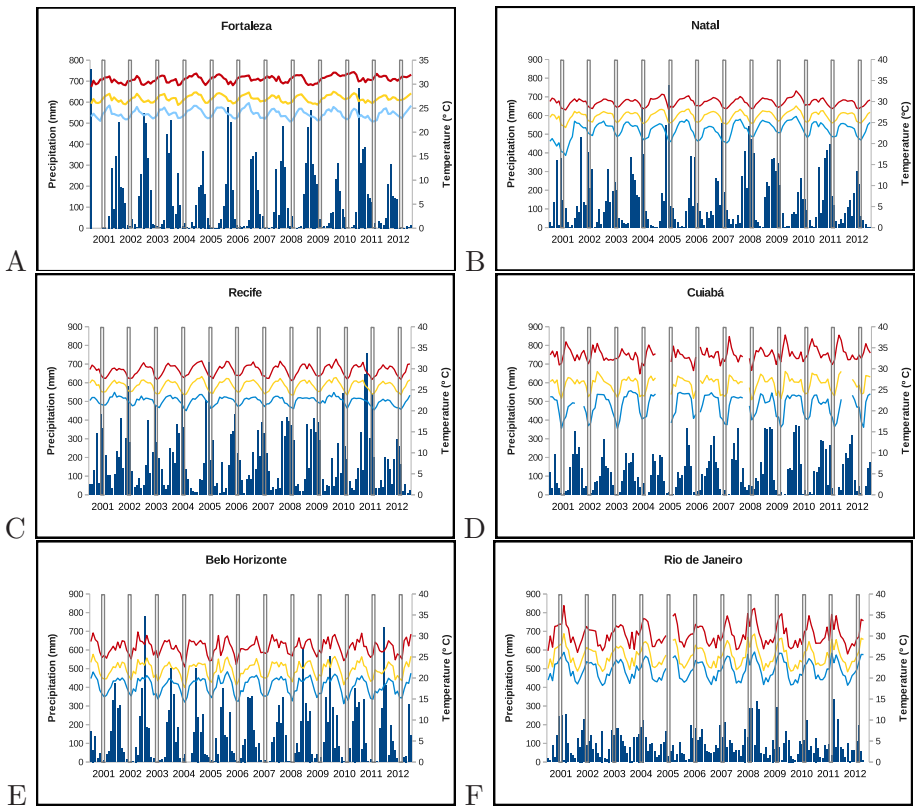


Figure 3: Time series of precipitation and temperature, from 2001 to 2012, for six of the Brazilian cities hosting the football games during the World Cup 2014. The white bars confine the the period where the World Cup is scheduled to happen, from 12 of June to 13 of July.

Exceptions on the described paten, where the peak of the dengue epidemic occur during the winter period: Fortaleza in 2005 and 2006 and in Recife in 2010, although with relatively low number of infections, less than 2 per 100 thousands individuals.

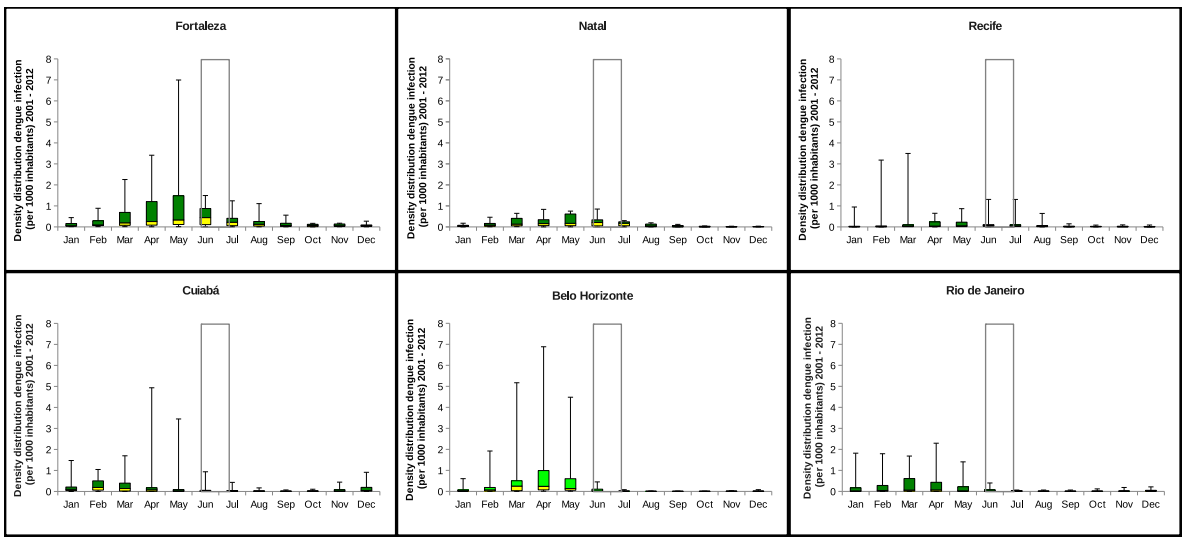


Figure 4: Box-plot for the density dengue infection in some of the Brazilian cities that are hosting the games. The white bars confine the the period where the World Cup is scheduled to happen, from 12 of June to 13 of July.

2.2 Box-plot Analysis

Using The box-plot, we present the minimum value, the 25%, 50% (median) and 75% quartiles and the maximum of density dengue infection for the Brazilian cities that are going to host the football games. The data is not normally distributed and therefore, we look at the median and quartiles instead of the mean and standard deviation, differently from what Simon Hay have done in [1].

Cities with frequent high outbreaks appears with a large median and third quartile, as opposed to cities which have not significant outbreaks, represented by the maximum bar only. Fortaleza and Belo Horizonte are the cities where large outbreaks have been reported more frequently. For all other cities, we observe that the median is very low, showing that the reported outbreaks during the 12 years period have been mild in term od number of cases.

In fig.4 we observe that during the period of the world Cup, Fortaleza appears to have the higher density of cases, among the other cities in study.

3 Conclusions

The results of Simon Hay published in [1] were obtained after analyzing the data with mean and standard deviation only, assuming that the data is normally distributed. The population density was also not taking into consideration, giving the false impression that Fortaleza, Natal and Salvador cities would have lots of cases happening during the winter, causing fear for the World Cup 2014 and promoting expensive accommodations with screened windows and doors and air conditioning required, for example.

In this manuscript we performed a more careful data analysis for the Brazilians cities that are hosting the games and we showed that the risk of being infected by a dengue virus is seasonal and also proportional to the population density in Brazil, increasing during the rainy season and the presence of vector and human population density. Despite some exceptions, the dengue season happens during the same period of every year and during the winter period, the number of cases are residual, bringing no risk for the football fans which are coming to Brazil for the World Cup 2014.

Those findings have important implications for the effectiveness of intervention measures that will be provided for Public Health Authorities for dengue control.

Acknowledgements

This work has been supported by the European Union under FP7 in the project DENFREE.

References

[1] Simon Hay Football fever could be a dose of dengue. *Nature*; **503:** 439 (2013).

[2] Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature*; **496** 504-507 (2013).

[3] Aguiar M. *et. al.* How much complexity is needed to describe the fluctuations observed in dengue hemorrhagic fever incidence data? *Ecological Complexity*; **16** 31-40 (2013).

[4] Aguiar M. *et. al.* Are we modelling the correct dataset? Minimizing false predictions for dengue fever in Thailand *Epidemiology & Infection Journal*; 1-13 (2014).

[5] Pan American Health Organization (PAHO). Dengue Regional Information: Number of Cases Retrieved from $http : //www.paho.org/hq/index.php?option = com_content\&view = article\&id = 264\&Itemid = 363\&lang = en$

[6] Jos Luis San Martìn *et al.* The Epidemiology of Dengue in the Americas Over the Last Three Decades: A Worrisome Reality. *Am. J. Trop. Med. Hyg.*; **82** 128–135 (2010).

 ISBN: 978-84-616-9216-3

[7] Ministry of Health Brazil. Sistema de Informação de Agravos de Notificação - SINAN Retrieved from $http://dtr2004.saude.gov.br/sinanweb/$.

[8] National Institute of Meteorology. Banco de Dados Meteorológicos para Ensino e Pesquisa - BDMEP Retrieved from $http://www.inmet.gov.br/projetos/rede/pesquisa/$.

[9] Vasconcelos P. F. C. et al. Epidemia de dengue em Fortaleza, Cear: inquérito soroepidemiológico aleatório. *Journal of Public Health*; **32** 447–454 (1998).

[10] Filipe Rocha *et al.* Carnival or football, is there a real risk for acquiring dengue fever in Brazil during holidays seasons? *in preparation.*

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Borda-type algorithms to aggregate partial rankings

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

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

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

emails: juanangel.aledo@uclm.es, jose.gamez@uclm.es,
d.molina@estudiante.uam.es

Abstract

In this work we study the rank aggregation problem in a general setting, that is, we approach the problem for any kind of ranking: complete or incomplete and with or without ties. The underlying idea behind our approach is to take into account the so-called extension set of a ranking, that is, the set of permutations that are *compatible* with the given ranking. Moreover, we propose two new *distances* to compare any kind of rankings. As an application of our proposal, we develop a hill climbing algorithm associated to our distances to deal with the rank aggregation problem.

Key words: rank aggregation problem, Kemeny ranking problem, Kendall distance, Borda method, Extension set, permutation, partial ranking, hill climbing algorithm
MSC 2000: 68R05, 05A99, 68W40

1 Introduction

Dealing with rankings is currently a hot topic in statistics and machine learning research. Perhaps the reason is the increasing availability of problems whose basic data are rankings (e.g. recommender systems, combinatorial optimization, preferences, etc.).

Rankings are a natural way to express preferences. Specifically, given a set of items $S^n = \{1, 2, \ldots, n\}$, a ranking π is an order of preference over (some of) these items. The case that has received more attention in the literature is the one in which all the items are ranked, that is, rankings are permutations of n elements [4]. However, real world problems usually deal with *incomplete* rankings, i.e. only p items are ranked, $2 \leq p < n$. This is the

case of users expressing preferences about a set of movies, books, etc., when they have no opinion about some items. Moreover, in these cases is also usual to use a preference system which allows the user to assign the same score to different items, so obtaining (incomplete) rankings with ties.

An important problem when dealing with rankings is to obtain a *consensus* ranking which best represents (summarizes) a given set of rankings. In the case of complete rankings, i.e. permutations, this problem is known as the *Kemeny ranking problem* [3, 5]. In the more general case when dealing with incomplete rankings (with or without ties), this problem is known as the *rank aggregation problem* [7] and has application in many real-world problems. In all the cases it is an NP-hard problem and its solution is a permutation (complete without ties ranking) . For this reason, it is usual to approach this problem by using greedy algorithms. Among them, the Borda algorithm is undoubtedly the preferred one, because of its good trade-off between efficiency and accuracy [6, 5].

Given a dataset of rankings, the Borda algorithm assigns points to the items according to their positions in the rankings of the dataset (the more preferred an item is, the more points it gets) and finally it computes the consensus ranking by ordering the items from the most valued to the less valued one. For permutations, the points are assigned easily by giving n points to the first ranked item, $n-1$ to the second one, and so on. The problem appears when dealing with arbitrary (incomplete with/without ties) rankings. In this sense, two different approaches can be followed: (i) to ignore, when assigning points, those items not included in the incomplete ranking (as the Modified Borda algorithm does [8, 7]), or (ii) to deal with the uncertainty associated to the items non appearing in a given ranking, that is, taking into account the positions of the ranking in which they could be placed. Our proposal belongs to the second approach, and particularly uses the concept of extension sets to manage the unobserved information (see [9, 10] for related research on this idea). Moreover, to evaluate how good the obtained consensus ranking is, we introduce a similarity measure based on an extension of the well-known Kendall tau distance that allows to compute the *distance* between any two arbitrary rankings.

Thus, we provide a new Borda-type method based on a normative approach to deal with the rank aggregation problem for *arbitrary* (incomplete with or without ties) rankings. As mentioned above, the idea is based on the use of extension sets, and particularly on its number of elements. We provide the mathematical expressions to efficiently compute the cardinality of the required extension sets, avoiding in this way a brute-force approach. Our proposal generalizes previous developments designed to cope only with particular types of rankings (i.e. incomplete or with ties, but not both types simultaneously). To test the goodness of the proposed algorithm we conduct an experimental study to compare it with a generalized version of the classical Borda algorithm that allows to consider arbitrary rankings.

Another interesting topic is, given two rankings, to establish a way to measure how

 ISBN: 978-84-616-9216-3

similar they are, that is, to compare them. In the case of complete rankings, i.e. permutations, many tools have been proposed to deal with this problem. Distances as *Kendall tau distance*, which measures the total number of pairwise inversions, and *Spearman's footrule distance*, which measures the l_1 distance between ranks, are those that have received more attention in the literature. When dealing with incomplete rankings or rankings with ties, some approaches have been proposed based on the disagreements that the rankings considered present. However, all these proposals compare only one type of rankings, and do not join every type of rankings together in the comparisons.

In this work we provide a new method for comparing *arbitrary* (incomplete with or without ties) rankings based on the use of the extension sets to take into account all the information that the rankings provide. Our proposal generalizes previous developments designed to cope only with particular types of rankings (i.e. incomplete or with ties, but not both types simultaneously). As an application of our proposal, we develop a hill climbing algorithm associated to our distances to deal with the rank aggregation problem.

 ISBN: 978-84-616-9216-3

References

[1] C. W. Misner, K. S. Thorne and J. A. Wheeler, *Gravitation*, Freeman, San Francisco, 1970.

[2] E. Witten, *Supersymmetry and Morse theory*, J. Diff. Geom. **17** (1982) 661–692.

[3] J. L. Kemeny and J. G. Snell *Mathematical models in the social sciences* Blaisdell, New York (1962).

[4] M. A. Fligner and J. S. Verducci *Distance based ranking models* Journal of the Royal Statistical Society **48**(3) (1986).

[5] A. Ali and M. Meila *Experiments with kemeny ranking: What works when?* Mathematical Social Sciences **64**(1) 28–40 (2012).

[6] J. Borda *Memoire sur les elections au scrutin* Histoire de l'Academie Royal des Sciences.

[7] F. Schalekamp and A. van Zuylen *Rank aggregation: Together we're strong* ALENEX. 38–51 (2009).

[8] P. Emerson *The original borda count and partial voting* Social Choice and Welfare **40**(2) 353–358 (2013).

[9] W. Cheng, J. Hühn and E. Hüllermeier *Decision tree and instance-based learning for label ranking* Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09, ACM 161–168 (2009).

[10] P. Kidwell, G. Lebanon and W. S. Cleveland *Visualizing incomplete and partially ranked data* IEEE Trans. Vis. Comput. Graph. **14**(6) 1356–1363 (2008).

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

An extension of parallel dynamical systems over graphs

Juan A. Aledo[1], Silvia Martinez[1] and Jose C. Valverde[1]

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

emails: JuanA.Aledo@uclm.es, Silvia.MSanahuja@uclm.es, Jose.Valverde@uclm.es

Abstract

In this work we introduce a wide generalization of dynamical systems over graphs, by considering that the states of the entities can take values in an arbitrary Boolean algebra with 2^p elements, $p \in \mathbb{N}$, $p \geq 1$. Then the orbit structure of these more general parallel dynamical systems over undirected graphs where the evolution operator is an arbitrary maxterm or minterm is analyzed. Finally, we also study the cases of parallel dynamical systems whose evolution update is defined by means of independent local Boolean functions.

Key words: Graph dynamical systems; orbit structure; Boolean algebras; Boolean functions.
MSC 2000: 37B99; 37E15; 37N99; 68R10; 94C10

1 Introduction

A *graph dynamical system* (GDS) is a dynamical system constructed over a graph whose vertices, named *entities*, can have different states, such that all these states together at a given time constitute a state of the system which can evolve thanks to an updating scheme. The states of the vertices are commonly modeled by the Boolean values 0 and 1, while the updating scheme consists of as many local functions as vertices and a series of rules that indicate the order in which the local functions act.

When all the local functions act synchronously the system is called *parallel* (PDS) [3, 4, 5, 6, 7, 12]. In contrast, when the local functions follow an order to act, the system is called *sequential* (SDS) [12, 20].

In the specific literature, other related topics appeared previously, as *cellular automata* (CA) [16, 22, 25, 26] and *Boolean networks* (BN) [17, 18], which are, in fact, particular cases of GDS.[1].

CA, when finite, can be considered as a special kind of PDS by considering cells as entities. Nevertheless, CA are restricted cases of PDS in several ways. First of all, for a CA seen as a PDS, the dependency graph, which is derived from the lattice and the neighborhood structure, is regular, whereas the graph of a general PDS is arbitrary. Secondly, CA have a fixed local function or rule, associated to every cell, while general PDS can have distinct local functions to update different entities, which can be the restriction of a global one (see [3, 12]) or independently defined (see [5]). Thus, general PDS can have more involved update schemes.

CA are also updated in a parallel or synchronous manner by applying local functions on a subset that contains the (state value of the) cell. Nevertheless, in the last few years some extensions of the concept of CA, considering sequential or asynchronous updating, have appeared in the literature (see [13, 19, 23]). In fact, the concept of SDS constitutes a generalization of such a CA extension.

BN are a generalization of (finite) Boolean CA but, at the same time, a particular case of GDS by considering nodes as entities. One of the main differences with CA is that, in BN, the state of each node is not affected necessarily by its neighbors, but potentially by any node in the network. Thus, the uniform structure of neighborhood in CA disappears. However, some homogeneity remains, since each node is affected by k connections with other (or the same) entities. This homogeneity makes BN a particular case of GDS, since in GDS connections can be totally arbitrary. Another important difference between BN and CA is that, for BN, local Boolean functions of $k-$variables are generated randomly, which provides a different update schedule for each entity. This idea has been carried out and extended for PDS in two directions. Firstly, as can be seen in [3], local Boolean functions acting on each entity can have different number of variables (what cannot occur for BN); and secondly, they can be totally independent for each entity [5].

GDS, as a concept that generalizes the aforementioned ones, is relatively young and unexplored. In fact, the first ideas appeared in [8], which constituted an important step in the development of the mathematical foundations for the theory of Computation. In this work, sequentially updated cellular automata (SCA) over arbitrary graphs are employed as a paradigmatic framework. This first work was followed by [9], [10] and [11], where the authors developed this theory, analyzing the asymptotic behavior of such mathematical models. Later, many other works have appeared in order to describe the behavior of these dynamical systems (see [3, 4, 5, 6, 7, 12]) and also as applications of them to other questions (see [14, 15]).

[1]The abbreviations GDS, PDS, SDS, CA and BN will be used for the singular and plural forms of the corresponding terms, since it seems better from an aesthetic point of view.

 ISBN: 978-84-616-9216-3

In all of these works, the entities in the model can only have two state values, i.e., each entity can be either activated or deactivated. This is usually modeled by means of Boolean variables $x_i \in \{0,1\}$, $i = 1, 2, \ldots, n$, where n is the number of entities, in such a way that $x_i = 1$ (resp. $x_i = 0$) means that the entity i is activated (resp. deactivated). However, the original definition of CA in [25] contemplates the possibility that the cells take state values in a finite set, although subsequently the majority of studies have been made in the case of Boolean CA. In fact, in experimental models, the state values of the entities can belong to a more general (finite) set. This situation naturally appears, for instance, when each entity can have different levels of activation or intensity, belonging to a totally ordered finite set which can be represented by $\{0, 1, \ldots, m\}$ (see [24] for this approach in the context of probabilistic Boolean networks); or when each entity consists of several sub-entities, which can be activated or deactivated.

This last conception has inspired our extended model in this work. In this sense, we introduce a wide generalization of GDS, by considering that the state values of the entities can belong to an arbitrary Boolean algebra B with 2^p elements, $p \in \mathbb{N}$, $p \geq 1$. This consideration widely extends the traditional one where it is assumed that every entity can take values only in the simplest Boolean algebra $\{0,1\}$.

In particular, we develop some techniques which allow us to study the orbit structure of these dynamical systems. As an application, we study the orbit structure of parallel dynamical systems over undirected graphs where the evolution operator is an arbitrary maxterm or minterm, using and generalizing at the same time the results in [3]. Moreover, taking into account the results in [5], we also analyze the case of parallel dynamical systems on general Boolean algebras whose evolution update scheme is defined by means of independent local functions chosen among $OR, AND, NAND$ and NOR. Finally, as a consequence, the results for parallel dynamical systems over directed dependency graphs in [4] and [5] can be also extended to this more general context.

 ISBN: 978-84-616-9216-3

Acknowledgements

This work has been partially supported by the Spanish national grant MTM2011-23221.

References

[1] C. W. Misner, K. S. Thorne and J. A. Wheeler, *Gravitation*, Freeman, San Francisco, 1970.

[2] E. Witten, *Supersymmetry and Morse theory*, J. Diff. Geom. **17** (1982) 661–692.

[3] J. A. Aledo, S. Martinez, F. L. Pelayo, J. C. Valverde, *Parallel Dynamical Systems on Maxterms and Minterms Boolean Functions*, Math. Comput. Model. **35** (2012) 666–671.

[4] J. A. Aledo, S. Martinez, J. C. Valverde, *Parallel dynamical systems over directed dependency graphs*, Appl. Math. Comput. **219** (2012) 1114–1119.

[5] J. A. Aledo, S. Martinez, J. C. Valverde, Parallel discrete dynamical systems on independent local functions, J. Comput. Appl. Math. **237** (2013) 335–339.

[6] J. A. Aledo, S. Martinez, J. C. Valverde, *Updating method for the computation of orbits in parallel and sequential dynamical systems*, Int. J. Comput. Math. **90(9)** (2013) 1796–1808.

[7] J. A. Aledo, S. Martinez, J. C. Valverde, *Parallel dynamical systems over special digraph classes*, Int. J. Comput. Math. **90(10)** (2013) 2039–2048

[8] C. L. Barret, C. M. Reidys, *Elements of a theory of computer simulation I*, Appl. Math. Comput. **98** (1999) 241–259.

[9] C. L. Barret, H. S. Mortveit, C. M. Reidys, *Elements of a theory of computer simulation II*, Appl. Math. Comput. **107** (2002) 121–136.

[10] C. L. Barret, H. S. Mortveit, C. M. Reidys, *Elements of a theory of computer simulation III*, Appl. Math. Comput. **122** (2002) 325–340.

[11] C. L. Barrett, H. S. Mortveit, C. M. Reidys, *ETS IV: sequential dynamical systems: fixed points, invertibility and equivalence*, Appl. Math. Comput. **134** (2003) 153–171.

[12] C. L. Barret, W. Y. C. Chen, M. J. Zheng, *Discrete dynamical systems on graphs and Boolean functions*, Math. Comput. Simul. **66** (2004) 487–497.

 ISBN: 978-84-616-9216-3

[13] H. J. Blok, B. Bergersen, *Synchronous versus asynchronous updating in the game of Life*, Phys. Rev. E **59** (1999) 3876–3879.

[14] C. Bisi, G. Chiaselotti, *A class of lattices and Boolean functions related to the Manickam-Miklos-Singhi conjecture*, Adv. Geom. **13** (2013) 1–27.

[15] G. Chiaselotti, G.Marino, P. A. Oliverio, D. Petrassi, *A discrete dynamical model of signed partitions*, J. Appl. Math. **1** (2013), Article ID 973501, 10 pages.

[16] J. Kari, *Theory of cellular automata: A survey*, Theoretical Computer Science **334** (2005) 3-33.

[17] S. A. Kauffman, *Metabolic stability and epigenesis in randomly constructed genetic nets*, J. Theor. Biol. **22** (1969) 437–467.

[18] S. A. Kauffman, *Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford, 1993.

[19] T. Mihaela, T. Matache, J. Heidel, *Asynchronous random Boolean network model based on elementary cellular automata rule 126*, Phys. Rev. E **71** (2005) 1–13.

[20] H. S. Mortveit, C. M. Reidys, *An Introduction to Sequential Dynamical Systems*, Springer, New York, 2007.

[21] K. Rosen, *Discrete Mathematics and Its Applications*, McGraw-Hill Education, 2011.

[22] J. L. Schiff, Cellular automata: a discrete view of the world, Wiley, New York, 2008.

[23] B. Schönfisch, A. de Roos, *Synchronous and asynchronous updating in cellular automata*, BioSystems **51** (1999) 123–143.

[24] R. V. Sole, B. Luque, S. A. Kauffman, *Phase transitions in random networks with multiple states*, Technical Report 00-02-011, Santa Fe Institute, 2000.

[25] S. Wolfram, *Statistical mechanics of cellular automata*, Rev. Mod. Phys. **55** (3) (1983) 601–644.

[26] S. Wolfram, *Cellular Automata and Complexity*, Addison-Wesley, New York, 1994.

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

A numerical study of a nonlocal degenerate parabolic problem

Rui M.P. Almeida[1], Stanislav N. Antontsev[2] and José C.M. Duque[1]

[1] *Department of Mathematics, Faculty of Science, University of Beira Interior*

[2] *Center for Mathematics and Fundamental Applications, Faculty of Science, University of Lisbon*

emails: ralmeida@ubi.pt, anton@ptmat.fc.ul.pt, jduque@ubi.pt

Abstract

The aim of this paper is to numerically study a class of nonlinear nonlocal degenerate parabolic equations. The convergence and error bounds are proved for a linearized Crank-Nicolson-Galerkin finite element method with polynomial approximations of degree $k \geq 1$. Some explicit solutions are obtained and used to test the implementation of the method in Matlab environment.

Key words: nonlocal, degenerate, parabolic, PDE

1 Introduction

In this work, we study parabolic problems with nonlocal nonlinearity of the following type

$$\begin{cases} u_t - \left(\int_\Omega u^2(x,t)dx \right)^\gamma \Delta u = f\,(x,t)\,, \quad (x,t) \in \Omega \times]0,T] \\ u\,(x,t) = 0\,, \quad (x,t) \in \partial\Omega \times]0,T] \\ u(x,0) = u_0(x)\,, \quad x \in \Omega \end{cases} \tag{1}$$

where Ω is a bounded open domain in $\mathbb{R}^d$, $d \geq 1$, $\gamma \geq \frac{1}{2}$ is a real constant, f and u_0 are continuous integrable functions.
These type of problems were studied initially by Chipot and Lovat in [6], where they proposed the equation

$$u_t - a(\int_\Omega u\ dx)\Delta u = f \tag{2}$$

 ISBN: 978-84-616-9216-3

to model the density of a population, for example of bacteria, subject to spreading. In this paper, the authors prove the existence and uniqueness of weak solutions. Equation (2) can also appear in the study of heat propagation or in epidemic theory.
The existence, uniqueness, asymptotic behavior of weak and strong solutions of parabolic equations and systems with nonlocal diffusion terms have been widely studied in the last two decades (see, for example, [12, 8, 10] and their references).
The numerical analysis and simulation of such problems have been less studied. In [1], Ackleh and Ke proposed and made some numerical simulations with a finite difference scheme in one dimension and a finite volume discretization in two space dimensions to approximate the solutions of a nonlocal PDE. Bendahmane and Sepulveda [5], in 2009, investigated the propagation of an epidemic disease modelled by a system of three nonlocal partial differential equations (PDE), in a physical domain $\Omega \subset \mathbb{R}^n$ $(n = 1, 2, 3)$. They established the existence of discrete solutions to finite volume scheme and its convergence to the weak solution of the PDE. In [7] the authors proved the optimal order of convergence for a linearized Euler-Galerkin finite element method for a nonlocal system with absorbtion, and presented some numerical results. Almeida et al. [3],[4], established the convergence and error bounds of the fully discrete solutions for a class of nonlinear equations and for systems of reaction-diffusion nonlocal type with moving boundaries, using a linearized Crank-Nicolson-Galerkin finite element method with polynomial approximations of any degree. In [9], Robalo et al. obtained approximate numerical solutions for nonlocal reaction-diffusion systems of this type with a Matlab code based on the moving finite element method (MFEM) with high degree local approximations.
In this paper, we analyze a different diffusion term, dependent on the L_2-norm of the solution. In most of the previous papers, it is assumed that the diffusion term is bounded with $0 < m \leq a(s) \leq M < \infty$, $s \in \mathbb{R}$, and so the problem is always nondegenerate. Here, we study a case were the diffusion term could be zero or infinity. This work is concerned with the proof of the convergence of a total discrete solution using a Crank-Nicolson-Galekin finite element method and the use of this method to study the behaviour of the weak solutions. To the best of our knowledge, these results are new for nonlocal reaction-diffusion equations with this type of diffusion term.
The paper is organized as follows. In Section 2, we formulate the problem and the hypotheses on the data. In Section 3, we prove the convergence of the semidiscrete solution. Section 4 is devoted to the proof of the convergence to a fully discrete solution. In Section 5, we obtain some explicit solutions and we use then to simulate some examples in Section 6. Finally, in Section 7, we draw some conclusions.

 ISBN: 978-84-616-9216-3

2 Statement of the problem

Let Ω be a bounded open domain in $\mathbb{R}^d$, $d \geq 1$, with Lipschitz-continuous boundary $\partial\Omega$, and T an arbitrary positive finite instant. We consider the problem of finding the function $u(x,t)$ which satisfies the following conditions

$$\begin{cases} u_t - a(u)\Delta u = f\,(x,t)\,, \quad (x,t) \in \Omega\times]0,T] \\ u\,(x,t) = 0\,, \quad (x,t) \in \partial\Omega\times]0,T] \\ u(x,0) = u_0(x)\,, \quad x \in \Omega \end{cases} \tag{3}$$

where $a(u) = \left(\int_\Omega u^2(x,t)dx\right)^\gamma$ with $\gamma \geq \frac{1}{2}$ and f and u_0 are continuous integrable functions. If $\gamma = 0$, we have the heat equation which is widely known. For $\gamma > 0$, the problem could degenerate if there is an extinction phenomenon, and for $\gamma < 0$, if the extinction occurs, the problem becomes singular.
This problem was studied in [2], where the authors proved the existence of weak solutions for $t \in [0,T]$ and the existence of a positive instant t^* such that these solutions are unique and classical for $t \in [0,t^*]$. These conclusions follow mainly from the next Lemmas which are proved in [2].

Lemma 1. *Suppose that $\gamma > \frac{1}{2}$. If $\int_\Omega u_0\ dx > 0$, then there exists a $t^* > 0$ such that $a(u) \geq m > 0$ for $t \in [0,t^*]$ where u is a weak solution of problem (3).*

Lemma 2. *If*

$$0 < m \leq \int_\Omega v^2\ dx, \int_\Omega w^2\ dx \leq M < \infty$$

then

$$|a(v) - a(w)| \leq C\|v - w\|,$$

where C may depend on γ , m and M.

This Lemmas prove the nondegeneracy and the Lipschitz-continuity of the diffusion term and will be needed in the proofs of the following sections.
In [2] the asymptotic behaviour of the solutions as time increases, was also studied.
In what follows, let $(\cdot,\cdot)$ and $\|\cdot\|$ be, respectively, the inner product and the norm in $L_2(\Omega)$. The definition of a weak solution to this problem is as follows:

Definition 3 (Weak solution)**.** *We say that the function u is a weak solution of Problem (3) if*

$$u \in L_2(0,T;H_0^1(\Omega)), \frac{\partial u}{\partial t} \in L_2(0,T;L_2(\Omega)), \tag{4}$$

the equality

$$(u_t, w) + a(u)(\nabla u, \nabla w) = (f, w) \tag{5}$$

is valid for all $w \in H_0^1(\Omega)$ and $t \in]0,T[$, and

$$u(x,0) = u_0(x), \quad x \in \Omega. \tag{6}$$

 ISBN: 978-84-616-9216-3

Next, we present a Lemma which proves the Lipschitz-continuity of the diffusion term and which will be needed in the proofs of the following sections.

3 Space discretization

Let $\mathcal{T}_h$ denote a partition of Ω into disjoint simplexes T_i, $i = 1, \ldots, nt$ such that no vertex of any simplex lies in the interior or on the side of another simplex, and let $h = \max\{diam(T_i), i = 1, \ldots, nt\}$. Moreover, let S_h^k denote the continuous functions on the closure $\bar{\Omega}$ of Ω, which are polynomials of degree k in each simplex of $\mathcal{T}_h$ and which vanish on $\partial\Omega$, that is,

$$S_h^k = \{W \in C_0^0(\bar{\Omega}) | W_{|T_i} \text{ is a polynomial of degree } k \text{ for all } T_i \in \mathcal{T}_h\}.$$

If $\{\varphi_j\}_{j=1}^{np}$ is a basis for S_h^k, then we can represent every $W \in S_h^k$ as $W = \sum_{j=1}^{np} w_j \varphi_j$. Given a smooth function u on Ω which vanishes on $\partial\Omega$, we may define its interpolant, denoted by $I_h u$, as the function of S_h^k which coincides with u at the points $\{P_j\}_{j=1}^{np}$, that is, $I_h u = \sum_{j=1}^{np} u_j \varphi_j$.

Lemma 4 ([11]). *If $u \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$, then*

$$\|I_h u - u\| + h\|\nabla(I_h u - u)\| \leq C h^{k+1} \|u\|_{H^{k+1}}.$$

Definition 5 ([11] Ritz projection). *A function $\tilde{U} \in S_h^k$ is said to be the Ritz projection of $u \in H_0^1(\Omega)$ onto S_h^k if it satisfies*

$$(\nabla \tilde{U}, \nabla W) = (\nabla u, \nabla W), \quad \text{for all } W \in S_h^k.$$

Lemma 6 ([11]). *If $u \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$, then*

$$\|\tilde{U} - u\| + h\|\nabla(\tilde{U} - u)\| \leq C h^{k+1} \|u\|_{H^{k+1}},$$

where C does not depend on h or k.

The semidiscrete problem, based on Definition 3, consists in finding $U(x,t)$ belonging to S_h^k, for $t \geq 0$, such that for all $W \in S_h^k$ and $t \in]0, t^*[$:

$$\begin{cases} (U_t, W) + a(U)(\nabla U, \nabla W) = (f, W) \\ U(x,0) = I_h u_0 \end{cases}. \tag{7}$$

Theorem 7. *For $\gamma > 0$, if u is the solution of Problem (3) and U is a solution of (7), then*

$$\|U - u\| \leq C h^{k+1}, \quad t \in]0, t^*],$$

where C does not depend on h or k.

 ISBN: 978-84-616-9216-3

In virtue of Lemmas 1 and 2, the proof follows from classical arguments, and so we will only present the main steps.

Proof. Let us consider

$$\|U-u\| \leq \|U-\tilde{U}\| + \|\tilde{U}-u\| = \|\theta\| + \|\rho\|,$$

where $\tilde{U}$ is the Ritz projection of u. By Lemma 6

$$\|\rho\| \leq Ch^{k+1}\|u\|_{H^{k+1}(\Omega)} \tag{8}$$

and for θ we have that

$$(\theta_t, W) + a(U)(\nabla\theta, \nabla W) = ((u-\tilde{U})_t, W) + (a(u)-a(U))(\nabla u, \nabla W).$$

Choosing $W=\theta$, the Cauchy inequality and Lemma 1 imply that

$$\frac{1}{2}\frac{d}{dt}\|\theta\|^2 + m\|\nabla\theta\|^2 \leq \frac{1}{2}\|\rho_t\|^2 + \frac{1}{2}\|\theta\|^2 + m\|\nabla\theta\|^2 + C(a(u)-a(U))^2\|\nabla u\|^2.$$

By Lemma 2,

$$\frac{d}{dt}\|\theta\|^2 \leq C\|\theta\|^2 + C\|\rho\|^2 + \|\rho_t\|^2.$$

Gronwall's Lemma permits us to conclude that

$$\|\theta\|^2 \leq C\|\theta(x,0)\|^2 + C\int_0^{t^*} \|\rho\|^2 \, dt + \int_0^{t^*} \|\rho_t\|^2 \, dt,$$

where the elements of the right hand side are bounded as follows:

$$\|\theta(x,0)\|^2 \leq Ch^{2(k+1)}\|u\|^2_{H^{k+1}(\Omega)},$$

$$\int_0^{t^*} \|\rho\|^2 \, dt \leq Ch^{2(k+1)} \int_0^{t^*} \|u\|^2_{H^{k+1}(\Omega)} \, dt,$$

$$\int_0^{t^*} \|\rho_t\|^2 \, dt \leq Ch^{2(k+1)} \int_0^{t^*} \left\| \frac{\partial u}{\partial t} \right\|^2_{H^{k+1}(\Omega)} dt.$$

If we assume that u is sufficiently regular, then

$$\|\theta\|^2 \leq Ch^{2(k+1)},$$

and, adding the estimate in (8), the result is proved. □

 ISBN: 978-84-616-9216-3

4 Time discretization

For the time discretization, we choose a multistep linearization of the Crank-Nicolson method with an initial predictor-corrector scheme. Consider the partition $[0,t^*]=\cup_{j=1}^{ni}[t_{j-1},t_j]$ with $\delta=t_j-t_{j-1}$. Let us define

$$\bar{\partial}U_n=\frac{U_n-U_{n-1}}{\delta},\ \hat{U}_n=\frac{U_n+U_{n-1}}{2},\ \bar{U}_n=\frac{3}{2}U_{n-1}-\frac{1}{2}U_{n-2}\text{ and }f_{n-1/2}=f(x,\frac{t_n+t_{n-1}}{2}).$$

The fully discrete approximation $U_n(x)\approx u(x,t_n)$, $n=1,\dots,ni$, belonging to S_h^k, is obtained in such a way that $U_0=I_hu_0$, and for all $W\in S_h^k$,

$$(\frac{U_{0,1}-U_0}{\delta},W)+a(U_0)(\nabla\left(\frac{U_{1,0}+U_0}{2}\right),\nabla W)=(f_{1/2},W),\tag{9}$$

$$(\bar{\partial}U_1,W)+a(\frac{U_{1,0}+U_0}{2})(\nabla\hat{U}_1,\nabla W)=(f_{1/2},W),\tag{10}$$

$$(\bar{\partial}U_n,W)+a(\bar{U}_n)(\nabla\hat{U}_n,\nabla W)=(f_{n-1/2},W),\quad n=2,\dots,ni.\tag{11}$$

Theorem 8. *Assuming $\gamma>0$, if u is the solution of problem (3) and U_n is the fully discrete solution defined by (9)-(11), then*

$$\|U_n-u(x,t_n)\|\le C(h^{k+1}+\delta^2),\quad n=1,\dots,ni,$$

where C does not depend on h, k or δ.

Proof. We first establish the result for $n=1$. Considering $\theta_{1,0}=U_{1,0}-\tilde{U}_1$, $\hat{\theta}_{1,0}=\frac{\theta_{1,0}+\theta_0}{2}$ and $\overline{\partial}\theta_{1,0}=\frac{\theta_{1,0}-\theta_0}{\delta}$, we have

$$(\bar{\partial}\theta_{1,0},W)+a(U_0)(\nabla\hat{\theta}_{1,0},\nabla W)=(f_{1/2},W)-((u_t)_{1/2},W)-a(u_{1/2})(\nabla u_{1/2},\nabla W)$$

$$+((u_t)_{1/2}-\bar{\partial}\tilde{U}_1,W)+(a(u_{1/2})\nabla u_{1/2}-a(U_0)\nabla\hat{u}_1,\nabla W).$$

Setting $W=\hat{\theta}_{1,0}$, and using the Poincaré and Hölder inequalities, we obtain

$$\frac{1}{2}\bar{\partial}\|\theta_{1,0}\|^2+m\|\nabla\hat{\theta}_{1,0}\|^2\le C(\|(u_t)_{1/2}-\bar{\partial}\tilde{U}_1\|+\|\nabla(u_{1/2}-\hat{u}_1)\|+\|u_{1/2}-U_0\|)\|\nabla\hat{\theta}_{1,0}\|.$$

Furthermore,

$$\|(u_t)_{1/2}-\bar{\partial}\tilde{U}_1\|\le\|(u_t)_{1/2}-\overline{\partial}u_1\|+\|\overline{\partial}u_1-\overline{\partial}\tilde{U}_1\|\le C\delta^2+Ch^{k+1},$$

$$\left\|\nabla(u_{1/2}-\hat{u}_1)\right\|\le C\delta\int_{t_0}^{t_1}\|\nabla u_{tt}\|\ dt\le C\delta^2,$$

 ISBN: 978-84-616-9216-3

$$\|u_{1/2} - U_0\| \leq \|u_{1/2} - u_0\| + \|u_0 - U_0\| \leq C\delta + Ch^{k+1}.$$

Hence

$$\overline{\partial}\|\theta_{1,0}\|^2 \leq C(h^{k+1} + \delta)^2$$

and we have the estimate

$$\|\theta_{1,0}\|^2 \leq \|\theta_0\|^2 + C\delta(h^{k+1} + \delta)^2 \leq C(h^{2(k+1)} + \delta^3).$$

Repeating this process for the corrector equation (10), we then arrive at

$$\frac{1}{2}\bar{\partial}\|\theta_1\|^2 + m\|\nabla\hat{\theta}_1\|^2 \leq C(\|(u_t)_{1/2} - \bar{\partial}\tilde{U}_1\| + \|\nabla(u_{1/2} - \hat{u}_1)\| + \|u_{1/2} - \frac{U_{1,0} - U_0}{2}\|)\|\nabla\hat{\theta}_1\|.$$

Now

$$\|u_{1/2} - \frac{U_{1,0} - U_0}{2}\| \leq \|u_{1/2} - \hat{\tilde{U}}_1\| + \|\hat{\tilde{U}}_1 - \frac{U_{1,0} - U_0}{2}\| \leq \|u_{1/2} - \hat{\tilde{U}}_1\| + \frac{1}{2}\|\theta_{1,0}\| + \frac{1}{2}\|\theta_0\| \leq$$

$$\leq C(h^{k+1} + \delta^2) + Ch^{k+1} + C(h^{k+1} + \delta^{\frac{3}{2}}) \leq C(h^{k+1} + \delta^{\frac{3}{2}}),$$

and so, by Cauchy's inequality, we conclude that

$$\overline{\partial}\|\theta_1\|^2 \leq C(h^{2(k+1)} + \delta^3),$$

whence

$$\|\theta_1\|^2 \leq \|\theta_0\|^2 + C\delta(h^{2(k+1)} + \delta^3) \leq C(h^{2(k+1)} + \delta^4).$$

In order to prove the result for $n \geq 2$, we apply the same process to the equation in (11) and use the estimate

$$\|u_{n-1/2} - \overline{U}_n\| \leq \|u_{n-1/2} - \overline{u}_n\| + \|\overline{u}_n - \overline{U}_n\| \leq \|u_{n-1/2} - \overline{u}_n\| + \|\overline{\rho}_n\| + \|\overline{\theta}_n\| \leq$$

$$\leq C\delta^2 + Ch^{k+1} + C(\|\theta_{n-1}\| + \|\theta_{n-2}\|)$$

to prove that

$$\frac{1}{2}\bar{\partial}\|\theta_n\|^2 + m\|\nabla\hat{\theta}_n\|^2 \leq C(\|(u_t)_{n-1/2} - \bar{\partial}\tilde{U}_n\| + \|\nabla(u_{n-1/2} - \hat{u}_n)\| + \|u_{n-1/2} - \bar{U}_n\|)\|\nabla\hat{\theta}_n\|,$$

and

$$\overline{\partial}\|\theta_n\|^2 \leq C\|\theta_{n-1}\|^2 + C\|\theta_{n-2}\|^2 + C(h^{(k+1)} + \delta^2)^2.$$

Iterating, we obtain

$$\|\theta_n\|^2 \leq (1+C\delta)\|\theta_{n-1}\|^2 + C\delta\|\theta_{n-2}\|^2 + C\delta(h^{k+1}+\delta^2)^2 \leq C\|\theta_1\|^2 + C\delta\|\theta_0\|^2 + C\delta(h^{k+1}+\delta^2)^2$$

and recalling the estimates for $\|\theta_0\|$, $\|\theta_1\|$ and $\|\rho_n\|$, the proof is complete. □

 ISBN: 978-84-616-9216-3

5 Explicit solution

In order to test the implementation of the discrete solution in a programming language, we need to find an explicit exact solution to the problem. We seek an explicit solution of the form

$$u(x,t) = k(x)l(t). \tag{12}$$

Then the first equation in (3) becomes

$$k(x)l'(t) - l^{2\gamma+1}(t)\left(\int_\Omega k^2(x)\ dx\right)^\gamma \Delta k(x) = f(x,t). \tag{13}$$

If l is chosen such that

$$l'(t) = -l^{2\gamma+1}(t) \Leftrightarrow l(t) = (2\gamma t - 2\gamma C)^{-\frac{1}{2\gamma}},\ \gamma \neq 0 \quad \text{or} \quad l(t) = Ce^{-t},\ \gamma = 0,\ C \in \mathbb{R} \tag{14}$$

then (13) has the form

$$k(x) + \left(\int_\Omega k^2(x)\ dx\right)^\gamma \Delta k(x) = \frac{f(x,t)}{-l^{2\gamma+1}(t)}. \tag{15}$$

To obtain a function $k(x)$ which only depends of x, we must assume that

$$\frac{f(x,t)}{-l^{2\gamma+1}(t)} = g(x) \Leftrightarrow f(x,t) = -g(x)l^{2\gamma+1}(t).$$

In this case, let $w(x,\alpha)$ be such that

$$w(x) + \alpha\Delta w(x) = g(x). \tag{16}$$

Then

$$k(x) = w(x, \left(\int_\Omega w^2\ dx\right)^\gamma) \tag{17}$$

is a solution of (15). But (17) is defined in an implicit way, and so in order to obtain k in an explicit form, we must solve the equation

$$\alpha = \left(\int_\Omega w^2(x,\alpha)\ dx\right)^\gamma. \tag{18}$$

Collecting (17), (14) and (12), we obtain an explicit solution for the first equation in (3). For $d = 1$, the equation in (16) becomes $w(x) + \alpha w''(x) = g(x)$, and if g is continuous in Ω, then it admits the solution

$$w(x) = C_1 sin(\frac{x}{\sqrt{\alpha}}) + C_2 cos(\frac{x}{\sqrt{\alpha}}) - \frac{1}{\sqrt{\alpha}}\int_0^x g(\xi) sin(\frac{\xi - x}{\sqrt{\alpha}})\ d\xi. \tag{19}$$

Remark 9. *The constants C, C_1 and C_2 must be chosen in such a way that u satisfies the initial data and boundary conditions.*

Remark 10. *The derivation of conditions for the solvability of the equation in (18) is under study .*

 ISBN: 978-84-616-9216-3

6 Numerical simulations

6.1 Example 1

Consider Problem (3) with $\gamma = 0.5$, $f = \frac{x^2}{(t+1)^2}$ and

$$u_0 = \frac{1 - 2\alpha + 2\alpha\cos(\frac{1}{\sqrt{\alpha}})}{\sin(\frac{1}{\sqrt{\alpha}})}\sin(\frac{x}{\sqrt{\alpha}}) - 2\alpha\cos(\frac{x}{\sqrt{\alpha}}) - x^2 + 2\alpha$$

with $\alpha = 0.223688785954835$. The solution is

$$u(x,t) = \left(\frac{1 - 2\alpha + 2\alpha\cos(\frac{1}{\sqrt{\alpha}})}{\sin(\frac{1}{\sqrt{\alpha}})}\sin(\frac{x}{\sqrt{\alpha}}) - 2\alpha\cos(\frac{x}{\sqrt{\alpha}}) - x^2 + 2\alpha\right)(t+1)^{-1}$$

In Figure 1, we show the obtained solution for $h = 10^{-2}$, $\delta = 10^{-3}$ and $k = 2$. As expected, [2] there is a decay of the solution as the time increases. We simulated the problem with different combinations of h, δ and k. The results are shown in Figures 2 and 3. In Figure 2, it is evident that the convergence of h is of order $k+1$, and, in Figure 3, we can observe that the error for δ is of order 2.

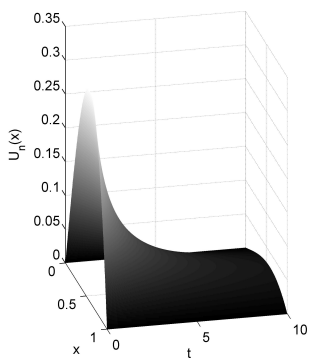


Figure 1: Evolution in time of the obtained solution in example 1 for $h = 10^{-2}$, $\delta = 10^{-3}$ and $k = 2$.

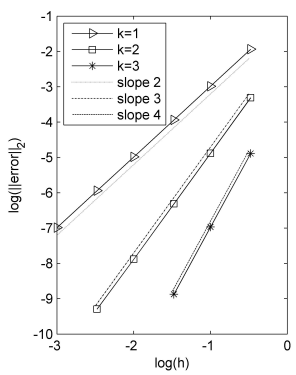


Figure 2: Study of convergence of h.

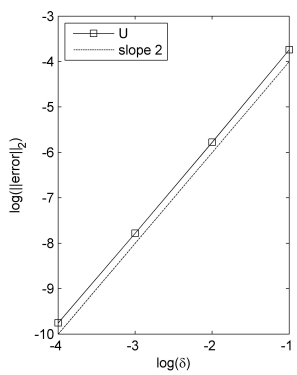


Figure 3: Study of convergence of δ.

 ISBN: 978-84-616-9216-3

6.2 Example 2

For $\gamma < 0$, we do not have the proof of the convergence, but we simulated an example. Consider Problem (3) with $\gamma = -0.5$, $f = 0$ and $u_0 = \sqrt{2}\pi^2 \sin(\pi x)$. From the last section, we have that $u(x,t) = \sqrt{2}\pi^2 \sin(\pi x)(1-t)$ is an exact explicit solution for Problem (3).

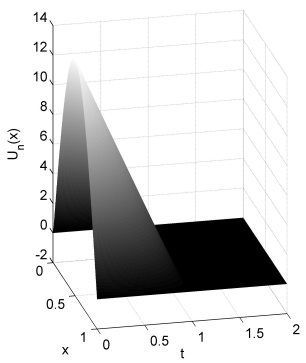


Figure 4: Evolution in time of the obtained solution in the example 2 for $h = 10^{-2}$, $\delta = 10^{-3}$ and $k = 2$.

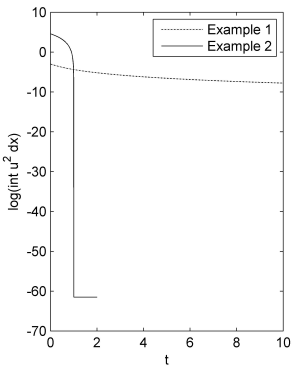


Figure 5: Study of the asymptotic behaviour.

In Figure 4, we show the obtained solution for $h = 10^{-2}$, $\delta = 10^{-3}$ and $k = 2$. As expected, we can observe an extinction in $t = 1$. The extinction is more evident in Figure 5, where we plotted the logarithm of the energetic function $y = \int_\Omega u^2 \, dx$ as a function of t, for Examples 1 and 2.

7 Conclusions

We proved optimal rates of convergence for a linearized Crank-Nicolson-Galerkin finite element method with piecewise polynomial of arbitrary degree basis functions in space when applied to a degenerate nonlocal parabolic equation. Some numerical experiments were presented, considering different functions f and exponent γ. The numerical results agree with the exact explicit solutions deduced and are in accordance with the theoretical results.

Acknowledgements

This work was partially supported by the research projects:
OE/MAT/UI0212/2011 - financed by FEDER through the - Programa Operacional Factores de Competitividade, FCT - Fundação para a Ciência e a Tecnologia and MTM2011-26119, MICINN, spain.

References

[1] Azmy S. Ackleh and Lan Ke. Existence-uniqueness and long time behavior for a class of nonlocal nonlinear parabolic evolution equations. *Proc. Amer. Math. Soc.*, **128**(12):3483–3492 (electronic), 2000.

[2] Rui M. P. Almeida, Stanislav N. Antontsev, and José C. M. Duque. On a nonlocal degenerate parabolic problem. Available from: http://arxiv.org/abs/, 2014.

[3] Rui M. P. Almeida, José C. M. Duque, Jorge Ferreira, and Rui J. Robalo. The Crank-Nicolson-Galerkin finite element method for a nonlocal parabolic equation with moving boundaries. Available from: http://ptmat.fc.ul.pt/arquivo/docs/preprints/pdf/2013/Jorge_Ferreira_Almeida_Duque_preprint_017_2013.pdf, 2013.

[4] Rui M. P. Almeida, José C. M. Duque, Jorge Ferreira, and Rui J. Robalo. Convergence of the Crank-Nicolson-Galerkin finite element method for a class of nonlocal parabolic systems with moving boundaries. Available from: http://arxiv.org/abs/1401.8220, 2014.

[5] Mostafa Bendahmane and Mauricio A. Sepúlveda. Convergence of a finite volume scheme for nonlocal reaction-diffusion systems modelling an epidemic disease. *Discrete Contin. Dyn. Syst. Ser. B*, **11**(4):823–853, 2009.

[6] M. Chipot and B. Lovat. Some remarks on nonlocal elliptic and parabolic problems. In *Proceedings of the Second World Congress of Nonlinear Analysts, Part 7 (Athens, 1996)*, **30**, pages 4619–4627, 1997.

[7] José C. M. Duque, Rui M. P. Almeida, Stanislav N. Antontsev, and Jorge Ferreira. The Euler-Galerkin finite element method for a nonlocal coupled system of reaction-diffusion type. Available from: http://ptmat.fc.ul.pt/arquivo/docs/preprints/pdf/2013/Duq_Ant_preprint_014_2013.pdf, 2013.

[8] José C. M. Duque, Rui M. P. Almeida, Stanislav N. Antontsev, and Jorge Ferreira. A reaction-diffusion model for the nonlinear coupled system: existence, uniqueness, long time behavior and localization properties of solutions. Available from: http://ptmat.fc.ul.pt/arquivo/docs/preprints/pdf/2013/preprint_2013_08_Antontsev.pdf, 2013.

[9] R. J. Robalo, R. M. Almeida, M. C. Coimbra, and J. Ferreira. Global solvability, exponential decay and MFEM approximate solution of a nonlinear coupled system with moving boundary. Available from: http://ptmat.fc.ul.pt/arquivo/docs/preprints/pdf/2013/preprint_015_CMAF_Jorge_Ferreira.pdf, 2013.

[10] R. J. Robalo, R. M. Almeida, M. C. Coimbra, and J. Ferreira. A reaction-diffusion model for a class of nonlinear parabolic equations with moving boundaries: existence, uniqueness, exponential decay and simulation. *Applied Mathematical Modelling*, 2014. Available from: http://dx.doi.org/10.1016/j.apm.2014.04.045.

[11] Vidar Thomée. *Galerkin finite element methods for parabolic problems*, **25** of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[12] S. Zheng and M. Chipot. Asymptotic behavior of solutions to nonlinear parabolic equations with nonlocal terms. *Asymptot. Anal.*, **45**(3-4):301–312, 2005.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

On the characterization of almost strictly totally negative matrices

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

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

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

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

Abstract

A real matrix $A = (a_{ij})_{1\leq i,j,\leq n}$ is said to be almost strictly totally negative if it is almost strictly sign regular with signature $\varepsilon = (-1,-1,\ldots,-1)$. In this paper a characterization of nonsingular almost strictly totally negative matrices is presented.

Key words: sign regular matrices, almost strictly totally negative matrices, Neville elimination
MSC 2000: 65F05, 15A48, 65F40

1 Introduction

Matrices with all its minors nonnegative, known as Totally Positive (TP) matrices, have been widely studied since mid of the last century (see e.g. [2–5]). They form a subclass of the set of the Sign Regular (SR) matrices, whose minors of the same order have the same sign (see e.g. [2,8]). Among the SR matrices, an important particular subclass is that of the Almost Strictly Sign Regular (ASSR) matrices, defined by R. Huang *et al.* [7] as that whose nontrivial minors of the same order have all the same strict sign. In this work the authors, that characterized this kind of matrices through the Neville Elimination (NE) procedure in a previous paper [1], deal with a subset of them called Almost Strictly Totally Negative (ASTN) matrices. All nontrivial minor of these matrices are strictly negative, which notably simplifies the characterization proposed in [1] for ASSR matrices.

The NE is an alternative procedure to Gaussian elimination for reducing a square matrix to upper triangular form, preferable for some classes of matrices and when using pivoting

 ISBN: 978-84-616-9216-3

strategies in parallel implementations. Roughly speaking, the Neville elimination introduces zeros in of each column of a matrix by adding to each row an appropriate multiple of the previous one (instead of using a single row with a fixed pivot, as in Gaussian elimination).

The ASSR matrices present grouped null elements in certain positions, and can be classified in two classes which are defined below, type-I and type-II staircase.

A matrix $A = (a_{ij})_{1\leq i,j\leq n}$ is called type-I staircase if it satisfies simultaneously the following conditions

- $a_{11} \neq 0,\ a_{22} \neq 0,\ \ldots, a_{nn} \neq 0$;
- $a_{ij} = 0,\ i > j \Rightarrow a_{kl} = 0,\ \forall l \leq j,\ i \leq k$;
- $a_{ij} = 0,\ i < j \Rightarrow a_{kl} = 0,\ \forall k \leq i,\ j \leq l$.

From now on it will be frequently used the backward identity matrix $n \times n$, P_n, whose element (i, j) is defined as

$$\begin{cases} 1, & \text{if } i + j = n + 1, \\ 0, & \text{otherwise.} \end{cases}$$

So, A is a type-II staircase matrix if it verifies that P_nA is a type-I staircase matrix.

To describe clearly the zero pattern of a nonsingular matrix A type-I staircase (or type-II staircase, using the $n \times n$ backward identity matrix P_n) we must introduce some notations. For a matrix $A = (a_{ij})_{1\leq i,j\leq n}$ type-I staircase, we define

$$i_0 = 1, \qquad j_0 = 1, \tag{1}$$

and for $k = 1, \ldots, l$:

$$i_k = \max\left\{i \ / \ a_{ij_{k-1}} \neq 0\right\} + 1 \ (\leq n + 1), \tag{2}$$

$$j_k = \max\left\{j \leq i_k \ / \ a_{i_k j} = 0\right\} + 1 \ (\leq n + 1), \tag{3}$$

where l is given in this recurrent definition by $j_l = n + 1$.

Analogously we define

$$\widehat{j}_0 = 1, \qquad \widehat{i}_0 = 1 \tag{4}$$

and for $k = 1, \ldots, r$:

$$\widehat{j}_k = \max\left\{j \ / \ a_{\widehat{i}_{k-1}j} \neq 0\right\} + 1 \ (\leq n + 1), \tag{5}$$

$$\widehat{i}_k = \max\left\{i \leq \widehat{j}_k \ / \ a_{i\widehat{j}_k} = 0\right\} + 1 \ (\leq n + 1), \tag{6}$$

where $\widehat{i}_r = n + 1$.

Finally, we denote by I, J, $\widehat{I}$ and $\widehat{J}$ the following sets of indices

$$\begin{array}{rclcrcl} I & = & \{i_0, i_1, \ldots, i_l\}, & & J & = & \{j_0, j_1, \ldots, j_l\}, \\ \widehat{I} & = & \left\{\widehat{i}_0, \widehat{i}_1, \ldots, \widehat{i}_r\right\}, & & \widehat{J} & = & \left\{\widehat{j}_0, \widehat{j}_1, \ldots, \widehat{j}_r\right\}, \end{array}$$

thereby defining the zero pattern in the matrix A.

 ISBN: 978-84-616-9216-3

2 Basic notations, definitions and auxiliary results

Since all the matrices of our concern are described through their minors, we are going to introduce some classic notations. For $m, n \in \mathbb{N}$, with $1 \leq m \leq n$, $Q_{m,n}$ denotes the set of all increasing sequences of m natural numbers not greater than n. For $\alpha = (\alpha_1, \ldots, \alpha_m)$, $\beta = (\beta_1, \ldots, \beta_m) \in Q_{m,n}$ and A an $n \times n$ real matrix, we denote $A[\alpha|\beta]$ the $m \times m$ submatrix of A containing rows $\alpha_1, \ldots, \alpha_m$ and columns $\beta_1, \ldots, \beta_m$ of A. If $\alpha = \beta$, we denote by $A[\alpha] := A[\alpha|\alpha]$ the corresponding principal minor. $Q^0_{m,n}$ denotes the set of increasing sequences of m consecutive natural numbers not greater than n.

Next, we present some definitions and basic results.

Definition 1 *For a real matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ type-I (type-II) staircase, a submatrix $A[\alpha|\beta]$, with $\alpha, \beta \in Q_{m,n}$ is nontrivial if all its main diagonal (secondary diagonal) elements are nonzero.*

The minor associated to a nontrivial submatrix ($A[\alpha|\beta]$) is called nontrivial minor ($\det A[\alpha|\beta]$).

Definition 2 *A vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is a signature sequence, or simply, a signature, if $|\varepsilon_i| = 1$, $\forall i \in \mathbb{N}$, $i \leq n$.*

Definition 3 *A real matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ is said to be ASSR with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ if all its nontrivial minors $\det A[\alpha|\beta]$ satisfy that*

$$\varepsilon_m \det A[\alpha|\beta] > 0, \quad \alpha, \beta \in Q_{m,n}, \quad m \leq n. \tag{7}$$

The backward identity matrix allows us relating the signatures of an ASSR matrix A and $P_n A$ by means of the following result (see [1]).

Proposition 1 *A real matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ is ASSR if and only if $P_n A$ it is also. Furthermore, if the signature of A is $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, then the signature of $P_n A$ is $\varepsilon' = (\varepsilon'_1, \varepsilon'_2, \ldots, \varepsilon'_n)$, with $\varepsilon'_m = (-1)^{\frac{m(m-1)}{2}} \varepsilon_m$.*

Rong Huang *et al.* prove in Theorem 10 of [7] the next characterization for ASSR matrices:

Theorem 1 *Let A be a real matrix $n \times n$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ be a signature. Then A is nonsingular ASSR with signature ε if and only if A is a type-I or type-II staircase matrix and all its nontrivial minors with $\alpha, \beta \in Q^0_{m,n}$, $m \leq n$, satisfy*

$$\varepsilon_m \det A[\alpha|\beta] > 0. \tag{8}$$

 ISBN: 978-84-616-9216-3

3 ASTN matrices

Below we present the definitions and results that allow us characterizing the ASTN matrices by means of the NE procedure.

Definition 4 *A real matrix $A = (a_{ij})_{1\leq i,j,\leq n}$ is said to be Totally Negative (TN) if all its minors satisfy*

$$\det A[\alpha|\beta] \leq 0, \quad \forall \alpha, \beta \in Q_{m,n}, \quad 1 \leq m \leq n. \tag{9}$$

Obviously a TN matriz is SR with signature $\varepsilon = (-1, -1, \ldots, -1)$.

Definition 5 *A real matrix $A = (a_{ij})_{1\leq i,j\leq n}$ is said to be Strictly Totally Negative (STN) if all its minors* $\det A[\alpha|\beta]$ *satisfy that*

$$\det A[\alpha|\beta] < 0, \quad \alpha, \beta \in Q_{m,n}, \quad m \leq n. \tag{10}$$

Definition 6 *A real matrix $A = (a_{ij})_{1\leq i,j,\leq n}$ is said to be Almost Strictly Totally Negative (ASTN) if it is ASSR with signature $\varepsilon = (-1, -1, \ldots, -1)$.*

Taking account that an ASTN matrix nonsingular is an ASSR matrix with $\varepsilon_2 = -1$, such matrix is type-II staircase. In [8] the author defines the sign regular matrices of order m (SR_m), as those satisfying for all $j = 1, \ldots, m$, that all their minors of order j have the same sign or are zero. By Theorem 2.1 of [8] we can describe the zero pattern for the ASTN matrices. This result assures that if $\varepsilon_1 = \varepsilon_3$ and $\varepsilon_2 = -1$, then $a_{ij} \neq 0$ for $(i, j) \in \{(1, 1), (n, n)\}$.

Remark 1 *Let $A = (a_{ij})_{1\leq i,j,\leq n}$ be a nonsingular ASTN matrix with $n \geq 2$. Then $B = P_n A = (b_{ij})_{1\leq i,j,\leq n}$ is a type-I staircase matrix and its zero pattern verifies $card(I), card(\widehat{I}) \leq 3$. In fact, its zero pattern only can be*

- *If $b_{ij} \neq 0$ for all i, j, then $I = [1, n+1]$, $J = [1, n+1]$, $\widehat{I} = [1, n+1]$, $\widehat{J} = [1, n+1]$.*
- *If b_{n1} is the only zero in B, then $I = [1, n, n+1]$, $J = [1, 2, n+1]$, $\widehat{I} = [1, n+1]$, $\widehat{J} = [1, n+1]$.*
- *If the only zero in B is b_{1n}, then $I = [1, n+1]$, $J = [1, n+1]$, $\widehat{I} = [1, 2, n+1]$, $\widehat{J} = [1, n, n+1]$.*
- *If the elements b_{1n} and b_{n1} are zero, then $I = [1, n, n+1]$, $J = [1, 2, n+1]$, $\widehat{I} = [1, 2, n+1]$, $\widehat{J} = [1, n, n+1]$.*

Next, a characterization of ASTN matrices is presented.

 ISBN: 978-84-616-9216-3

Theorem 2 *Given a nonsingular matrix A $n \times n$, with $n \geq 2$, A is ASTN if and only if the following properties hold simultaneously:*

(a) A has all its elements not zero, except at most those who occupy the positions $(1,1)$ and (n,n).

(b) The NE of $B = P_n A$ and $\widetilde{B} = P_n A^T$ can be performed without row exchanges.

(c) The pivots p_{ij} of the NE of B, with $i \geq j$ verify:

$$p_{n1} = 0 \Leftrightarrow b_{n1} = 0, \tag{11}$$

$$if\ j = j_t,\ then\ p_{ij} < 0 \Leftrightarrow b_{ij} \neq 0, \tag{12}$$

$$if\ j > j_t,\ then\ (-1)^{j-j_t} p_{ij} > 0 \Leftrightarrow b_{ij} \neq 0, \tag{13}$$

and the pivots q_{ij} of $\widetilde{B}$ with $i < j$ verify

$$q_{1n} = 0 \Leftrightarrow b_{1n} = 0, \tag{14}$$

$$if\ i = \widehat{i}_t,\ then\ q_{ij} < 0 \Leftrightarrow b_{ij} \neq 0, \tag{15}$$

$$if\ i > \widehat{i}_t,\ then\ (-1)^{i-\widehat{i}_t} q_{ij} > 0 \Leftrightarrow b_{ij} \neq 0, \tag{16}$$

where

$$j_t = \max\left\{j_s \in J\ /\ 0 \leq s \leq k-1,\ j - j_s \leq i - i_s\right\}, \tag{17}$$

$$\widehat{i}_t = \max\left\{\widehat{i}_s \in \widehat{I}\ /\ 0 \leq s \leq k'-1,\ i - \widehat{i}_s \leq j - \widehat{j}_s\right\}, \tag{18}$$

k is the only index satisfying $j_{k-1} \leq j < j_k$ and k' is the only index satisfying $\widehat{i}_{k'-1} \leq i < \widehat{i}_{k'}$.

(d) The matrix $M = A[1,\ldots,n-1|2,\ldots n]$ is STN.

If an ASTN matrix A verifies that $a_{11} \neq 0$ and $a_{nn} \neq 0$ then it is an STN matrix. By the Remark 3.6 of [6], it is possible to affirm that a nonsingular matrix A that verifies $a_{nn} < 0$ is STN if and only if the NE of A and A^T can be performed without row exchanges with positive multipliers and with diagonal pivots verifying

$$p_{11} < 0, \qquad p_{ii} > 0 \quad \forall i > 1. \tag{19}$$

By this way, if A is an STN matrix we can characterize it analogously with Theorem 2 supposing that it is an ASTN matrix without zero positions.

Theorem 3 *If A is a nonsingular matrix $n \times n$, $n \geq 2$, $a_{11} < 0$ and $a_{nn} < 0$. Then A is an STN matrix if and only if the following properties hold simultaneously:*

 ISBN: 978-84-616-9216-3

(a) The NE of $B = P_n A$ and $\widetilde{B} = P_n A^T$ can be performed without row exchanges.

(b) The pivots p_{ij} of the NE of B, with $i \geq j$ verify:

$$p_{i1} < 0, \tag{20}$$

$$j > 1, \text{ then } (-1)^{j-1} p_{ij} > 0, \tag{21}$$

and the pivots q_{ij} of $\widetilde{B}$ with $i < j$ verify

$$q_{1j} < 0, \tag{22}$$

$$i > 1, \text{ then } (-1)^{i-1} q_{ij} > 0. \tag{23}$$

Acknowledgements

This work has been partially supported by the Spanish Research Grant MTM2012-31544 and under MEC and FEDER Grant TEC2012-38142-C04-04.

References

[1] P. Alonso, J.M. Peña and M.L. Serrano, *On the Characterization of Almost strictly sign regular matrices*, J. Comput. Appl. Math. **http://dx.doi.org/10.1016/j.cam.2014.01.032** (2014).

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

[3] Colin W. Cryer, *Some Properties of Totally Positive Matrices*, Linear Algebra Appl. **15** (1976) 1–25.

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

[5] M. Gasca, J.M. Peña, *Total positivity and Neville elimination*, Linear Algebra Appl. **165** (1992) 25–44.

[6] M. Gasca, J.M. Peña, *A Test for Strictly Sign-Regularity*, Linear Algebra Appl. **198** (1994) 133–142.

[7] R. Huang, J. Liu, L. Zhu, *Nonsingular almost strictly sign regular matrices*, Linear Algebra Appl. **436** (2012) 4179–4192.

[8] J.M. Peña, *Sign Regular Matrices of Order Two*, Linear Algebra Appl. **50** (2002) 91–97.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
03–07 July, 2014.

High Dimensional Model Representation in Image Processing

Emine Mine Altın[1] and Burcu Tunga[1]

[1] *Department of Mathematical Engineering, Faculty of Science and Letters, Istanbul Technical University, , Maslak 34469, İstanbul, Turkey*

emails: altinemi@itu.edu.tr, tungab@itu.edu.tr

Abstract

This study deals with testing whether the High Dimensional Model Representation (HDMR) method can be used as an image reconstruction method by investigating the performance of the method in representing images. HDMR is a method which intents to decompose the given multivariate function and/or multivariate data. In this study, since a true color image object is a three-dimensional array which stores the color values for each pixel as RGB triplets in MATLAB, we use HDMR method to decompose this three-dimensional array. After the decomposition we obtain HDMR components and using superposition of these components we create new images. Finally, we try to determine the quality of these images. This study has also some illustrative applications.

Key words: HDMR, Image reconstruction, Decomposition, Image representation.

1 Introduction

Mostly, scientific studies including multivariate functions are stalled by the dimensionality of the problem. To overcome this and decrease dimension, scientists try to find new methods. High Dimensional Model Representation (HDMR) is one of these methods and is used for twenty years [1–5]. HDMR and its variaties allow us to make highly accurate approximations using less variate functions instead of the given multivariate function. HDMR method is used in two different ways in the literature. One of them is to decompose a multivariate function given in analytical form and the other one is to partition the given multivariate data when we only know the function values at the nodes of the problem domain. In this work, the second way is used to make image representation.

Image reconstruction is considered as generation of a image from scattered data set and it is useful in medicine, biology, earth science, archeology, materials science and astronomy

[6–9]. The aim of image reconstruction is to recover an image that has been degraded by some mathematical and statistical models [10, 11]. In this work, we use a three-dimensional array that represents an image. We try to partition that multivariate data set into less variate data sets and to create the image using these partitioned data sets. When that data set is once obtained through Image Processing Toolbox of MATLAB [12] as RGB format, we investigate at what level the HDMR method is successful to reconstruct the image as an matematical method. For this purpose, at first the univariate components of HDMR method are obtained and an image is produced by superpositioning these components. However, this image includes the color of the original image the characteristics of it are not reflected by the univariate terms. Hence, if the bivariate terms are evaluated and an image is created by considering these bivariate terms, it is seen that both the pattern and the color of the original image are obtained. Here, the basic issue is whether the given image is represented exactly. Our efforts show that if the original image is grayscale then the HDMR method can represent the given image exactly but the original image is colored one then the HDMR method can not represent the image. All these findings are given in fourth section. To represent the colored image exactly by using HDMR method, we need an improvement in the method but this is considered as further work. So, this work aims to investigate whether we can use HDMR as an image reconstruction method.

This paper is organized as follows. The second section includes some of main issues of the High Dimensional Model Representation (HDMR), how the HDMR method is applied to the image processing is given in the third section. The fourth section covers the numerical results and some images which are obtained through the HDMR method. The final section consists of certain concluding remarks.

2 Mathematical Background

2.1 High Dimensional Model Representation

The basic expansion formula of HDMR for a given multivariate function, $f(x_1, \ldots, x_N)$, is given as follows.

$$f(x_1, \ldots, x_N) = f_0 + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) + \sum_{\substack{i_1, i_2=1 \\ i_1<i_2}}^{N} f_{i_1 i_2}(x_{i_1}, x_{i_2}) + \cdots + f_{12\ldots N}(x_1, \ldots, x_N) \quad (1)$$

This is a finite expansion and it consists of a constant term f_0, univariate terms $f_{i_1}(x_{i_1})$, bivariate terms $f_{i_1 i_2}(x_{i_1}, x_{i_2})$ and higher variate terms. The critical issue in the HDMR algorithm is to determine the right hand side components of HDMR. For this purpose, we use the following vanishing conditions under some normalization conditions defined on

weight factors to uniquely obtain the components of the HDMR expansion

$$\int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N W(x_1, \ldots, x_N) f_i(x_i) = 0, \quad \int_{a_j}^{b_j} dx_j W_j(x_j) = 1 \tag{2}$$

where $1 \leq j \leq N$ and $W_j(x_j)$s are weight factors and the product of these factors constructs the weight function, $W(x_1, \ldots, x_N)$ [1,2].

$$W(x_1, \ldots, x_N) \equiv \prod_{j=1}^{N} W_j(x_j), \quad x_j \in [\, a_j \,,\, b_j \,], \quad 1 \leq j \leq N \tag{3}$$

The HDMR components are determined through multiple integrations by also taking these vanishing and normalization conditions into consideration. To obtain the general structure of the constant component, we need to apply an operator having N-tuple integrals with the weight function given in (3) to the both sides of the HDMR expansion. The following relation is obtained as the constant component structure after some calculations

$$f_0 = \int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N W(x_1, \ldots, x_N) f(x_1, \ldots, x_N). \tag{4}$$

In a similar manner the general structure of each univariate component can also be determined, however, this time we need to use an operator with $(N-1)$-tuple integrations. The independent variable discarded from the integration is the variable of the targeted univariate component. Finally, the following relation is obtained as the general structure of the univariate HDMR components

$$\begin{aligned} f_i(x_i) \quad \equiv \quad & \int_{a_1}^{b_1} dx_1 W_1(x_1) \cdots \int_{a_{i-1}}^{b_{i-1}} dx_{i-1} W_{i-1}(x_{i-1}) \int_{a_{i+1}}^{b_{i+1}} dx_{i+1} W_{i+1}(x_{i+1}) \\ & \times \cdots \int_{a_N}^{b_N} dx_N W_N(x_N) f(x_1, \ldots, x_N) - f_0, \quad 1 \leq i \leq N \end{aligned} \tag{5}$$

Higher variate terms can be found by using same philosophy [2].

Since the main aim of this study is to develop a new algorithm to be used in reconstruction of images. That is, we need to perform this new algorithm on finite number of data. Therefore, we have to rewrite above relations in terms of summations. For this reason, Dirac delta functions are utilized in the weight factors [3]

$$W_j(x_j) \equiv \sum_{k_j=1}^{n_j} \alpha_{k_j}^{(j)} \delta\left(x_j - \xi_j^{(k_j)}\right), \quad x_j \in [\, a_j \,,\, b_j \,], \quad 1 \leq j \leq N \tag{6}$$

where α parameters are used to give a different importance to each node of the data set that represents the image under consideration. To this end, the constant component is obtained

 ISBN: 978-84-616-9216-3

as follows

$$f_0 = \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \cdots \sum_{k_N=1}^{n_N} \left(\prod_{i=1}^{N} \alpha_{k_i}^{(i)} \right) f(\xi_1^{(k_1)}, \ldots, \xi_N^{(k_N)}) \tag{7}$$

The structure of the univariate terms can be obtained as follows when we use the weight function whose factors are composed of Dirac delta functions given in (6).

$$\begin{aligned} f_m\left(\xi_m^{(k_m)}\right) &= \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \cdots \sum_{k_{m-1}=1}^{n_{m-1}} \sum_{k_{m+1}=1}^{n_{m+1}} \cdots \sum_{k_N=1}^{n_N} \left(\prod_{\substack{i=1 \\ i \neq m}}^{N} \alpha_{k_i}^{(i)} \right) f(\xi_1^{(k_1)}, \ldots, \xi_N^{(k_N)}) - f_0 \\ & 1 \leq k_m \leq n_m, \quad 1 \leq m \leq N \end{aligned} \tag{8}$$

The bivariate components can be determined in the same manner and the following relation is obtained

$$\begin{aligned} f_{m_1 m_2}\left(\xi_{m_1}^{(k_{m_1})} \xi_{m_2}^{(k_{m_2})}\right) &= \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \cdots \sum_{k_{m_1-1}=1}^{n_{m_1-1}} \sum_{k_{m_1+1}=1}^{n_{m_1+1}} \cdots \sum_{k_{m_2-1}=1}^{n_{m_2-1}} \sum_{k_{m_2+1}=1}^{n_{m_2+1}} \cdots \\ & \times \sum_{k_N=1}^{n_N} \left(\prod_{\substack{i=1 \\ i \neq m_1 \wedge i \neq m_2}}^{N} \alpha_{k_i}^{(i)} \right) f(\xi_1^{(k_1)}, \ldots, \xi_N^{(k_N)}) - f_{m_1}\left(\xi_{m_1}^{(k_{m_1})}\right) \\ & - f_{m_2}\left(\xi_{m_2}^{(k_{m_2})}\right) - f_0 \end{aligned} \tag{9}$$

where $1 \leq k_{m_1} \leq n_{m_1}$, $1 \leq k_{m_2} \leq n_{m_2}$, and $1 \leq m_1, m_2 \leq N$.

Getting constant, univariate and bivariate HDMR components under Dirac delta type weight correspond to partitioning the given multivariate data into a constant value, univariate and bivariate data sets like below.

$$\begin{aligned} s_0(x_1, \ldots, x_N) &= f_0 \\ s_1(x_1, \ldots, x_N) &= s_0(x_1, \ldots, x_N) + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) \\ s_2(x_1, \ldots, x_N) &= s_1(x_1, \ldots, x_N) + \sum_{\substack{i_1, i_2=1 \\ i_1 < i_2}}^{N} f_{i_1 i_2}(x_{i_1}, x_{i_2}) \end{aligned} \tag{10}$$

where $s_0(x_1, \ldots, x_N)$, $s_1(x_1, \ldots, x_N)$ and $s_2(x_1, \ldots, x_N)$ are the consecutive summation of data sets. Since we are dealing with image reconstruction and we have three independent variables, at most the bivariate HDMR components can be used in the representation of the image through HDMR expansion.

3 Image Processing Through HDMR

In this section, we reconstruct the HDMR method to represent the image. A colored image has a 3-dimensional finite data set. So, there are 3 independent variables, x_1, x_2, x_3 and each variable takes on n_1, n_2 and n_3 number of different values respectively.

Hence, we can write the constant term of HDMR component as follows

$$f_0 = \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \sum_{k_3=1}^{3} \alpha_{k_1}^{(1)} \alpha_{k_2}^{(2)} \alpha_{k_3}^{(3)} f(\xi_1^{(k_1)}, \xi_2^{(k_2)}, \xi_3^{(k_3)}) \tag{11}$$

where $f(\xi_1, \xi_2, \xi_3)$ denotes the values of image in RGB format. Since RGB format is used here n_3 is always taken as 3. $\alpha_{k_1}^{(1)}$, $\alpha_{k_2}^{(2)}$ and $\alpha_{k_3}^{(3)}$ are named as weight factors and they are utilized to give different importance level to each datum.

If the normalization conditions, given in relation (2), on the weight factors are applied with the help of Dirac delta function [3], the following relation about α parameters is obtained.

$$\sum_{k_j=1}^{n_j} \alpha_{k_j}^{(j)} = 1, \quad 1 \leq j \leq 3 \tag{12}$$

In this work, for simplicity we choose the α parameter to be same for the related dimension. Under all these circumstances and using relation (8) given in previous section, we obtain the univariate components of HDMR.

$$\begin{aligned}
f_1\left(\xi_1^{(k_1)}\right) &= \sum_{k_2=1}^{n_2} \sum_{k_3=1}^{3} \alpha_{k_2}^{(2)} \alpha_{k_3}^{(3)} f(\xi_1^{(k_1)}, \xi_2^{(k_2)}, \xi_3^{(k_3)}) - f_0 \\
f_2\left(\xi_2^{(k_2)}\right) &= \sum_{k_1=1}^{n_1} \sum_{k_3=1}^{3} \alpha_{k_1}^{(1)} \alpha_{k_3}^{(3)} f(\xi_1^{(k_1)}, \xi_2^{(k_2)}, \xi_3^{(k_3)}) - f_0 \\
f_3\left(\xi_3^{(k_3)}\right) &= \sum_{k_1=1}^{n_1} \sum_{k_2=1}^{n_2} \alpha_{k_1}^{(1)} \alpha_{k_2}^{(2)} f(\xi_1^{(k_1)}, \xi_2^{(k_2)}, \xi_3^{(k_3)}) - f_0
\end{aligned} \tag{13}$$

Each univarite function demonstrates a two dimensional array that is corresponding to a matrix. The superposition of these three matrices and the constant term of HDMR, f_0, constructs the image under consideration. Our research shows that the obtained image is not well enough to represent the original image. Because the univariate HDMR approximant cannot represent its pattern that is, it is inefficient to achieve representation. So we need

 ISBN: 978-84-616-9216-3

to calculate bivariate components of HDMR as follows

$$\begin{aligned} f_{12}\left(\xi_1^{(k_{m_1})},\xi_2^{(k_{m_2})}\right) &= \sum_{k_3=1}^{3}\alpha_{k_3}^{(3)}f(\xi_1^{(k_1)},\xi_2^{(k_2)},\xi_3^{(k_3)}) - f_1\left(\xi_1^{(k_1)}\right) - f_2\left(\xi_2^{(k_2)}\right) - f_0 \\ f_{13}\left(\xi_1^{(k_{m_1})},\xi_3^{(k_{m_3})}\right) &= \sum_{k_2=1}^{n_2}\alpha_{k_2}^{(2)}f(\xi_1^{(k_1)},\xi_2^{(k_2)},\xi_3^{(k_3)}) - f_1\left(\xi_1^{(k_1)}\right) - f_3\left(\xi_3^{(k_3)}\right) - f_0 \\ f_{23}\left(\xi_2^{(k_{m_2})},\xi_3^{(k_{m_3})}\right) &= \sum_{k_1=1}^{n_1}\alpha_{k_1}^{(1)}f(\xi_1^{(k_1)},\xi_2^{(k_2)},\xi_3^{(k_3)}) - f_2\left(\xi_2^{(k_2)}\right) - f_3\left(\xi_3^{(k_3)}\right) - f_0 \end{aligned} \tag{14}$$

Each bivariate function is an one dimensional array and is obtained through the formulation (9). Since we have 3 independent varibles we evaluate at most bivarite terms of HDMR. In order to obtain image through HDMR method, the summation of constant, univariate and bivariate terms of HDMR should be used, that is, we have to evaluate second order HDMR approximant, s_2, which is given in (10) like below

$$s_2(x_1,x_2,x_3) = f_0 + \sum_{i_1=1}^{3} f_{i_1}(x_{i_1}) + \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{3} f_{i_1 i_2}(x_{i_1},x_{i_2}) \tag{15}$$

When the mentioned approximants are obtained through the HDMR method, a new problem comes out about how qualified these approximants and images constructed through either univarite or bivariate approximation in representing the original image. To answer this question, we can use some measures. In this work, we prefer to use relative error analysis to measure it. The related formulation is given as follows;

$$\mathcal{N}_{s_i} = \frac{\|f_{original} - f_{s_i}\|}{\|f_{original}\|} \tag{16}$$

where $f_{original}$ denotes the original image data set and f_{s_i}, $i = 1, 2$ is the new data set which is evaluated by using either first or second order HDMR aproximant respectively.

4 Findings

To determine the efficiency of the newly developed algorithm, we chose some images and we obtained a representation for each image by using HDMR method. To be able to achieve this, we wrote a code in MATLAB [12] environment. Hence we obtained image's pixel map as three-dimensional array which consists of three m by n matrices and then we applied HDMR method to this three-dimensional array to decompose that array. Hence we obtained

(a) Constant (b) Univariate (c) Bivariate (d) Original

Figure 1: Comparison of constant, univariate, bivariate approximants of HDMR and the original image

a constant value, some vectors and some matrices after decomposition. That is, a constant value, vectors and matrices correspond the constant, univariate and bivariate components of HDMR method respectively. To compose the HDMR approximants, we used the formula given in (10).

The testing images were selected from either well known images used in image processing literature or taken by the authors as photos from the real life. As mentioned before, first we got three-dimensional pixel map of each image. However, the first and second dimensions, which is $256x256x3$, are same in each image. The first two dimensions can be selected differently while the third one must remain 3 because of the RGB format

When an image is represented by HDMR, we can find some approximated images to the original one by evaluating s_0, s_1 and s_2 as given in (10). In Figure 1, we created three different images by using constant approximant s_0, univariate approximant s_1 and bivariate approximant s_2 respectively. The first figure of Figure 1 was produced by constant approximant including nothing about the image. For this reason we can not represent the original image by that. While the second one created by the univariate approximant has a colored pattern in it, it still represents nothing. The last image presented by bivariate approximant has enough information to show the image under consideration. Hence this approximant is very adequate to represent the image. In the other words, the image information is hidden in bivariate components. All these characteristics and performance differerences between the HDMR approximants can be easily seen in Figure 1.

When we apply HDMR method to grayscale image, the picture obtained from the bivariate approximant represents the original picture exactly. This is seen in Figure 2. On the other hand, if the relative error formula given in relation (16) is used, then the relative error for the given grayscale image is found as 0. It means that the grayscale image can be represented by bivariate HDMR approximant exactly. This result is shown in Table 1. This table also includes $\mathcal{N}_{s_1}$, the relative error of the obtained approximant with the constant and the univariate terms and $\mathcal{N}_{s_2}$, the relative error of the obtained approximant with the

 ISBN: 978-84-616-9216-3

(a) Tram (b) Representation

Figure 2: Comparison of the gray scale pictures with bivariate approximant of HDMR

Table 1: Relative Errors for different images representation through HDMR

	$\mathcal{N}_{Peppers}$	$\mathcal{N}_{Tram}$	$\mathcal{N}_{lena}$	$\mathcal{N}_{Sunflower}$	$\mathcal{N}_{Dog}$	$\mathcal{N}_{Girl}$
$\mathcal{N}_{s_1}$	0.3999	0.3437	0.2863	0.4862	0.2741	0,2792
$\mathcal{N}_{s_2}$	0.0632	0	0.0984	0.2174	0	0,0632

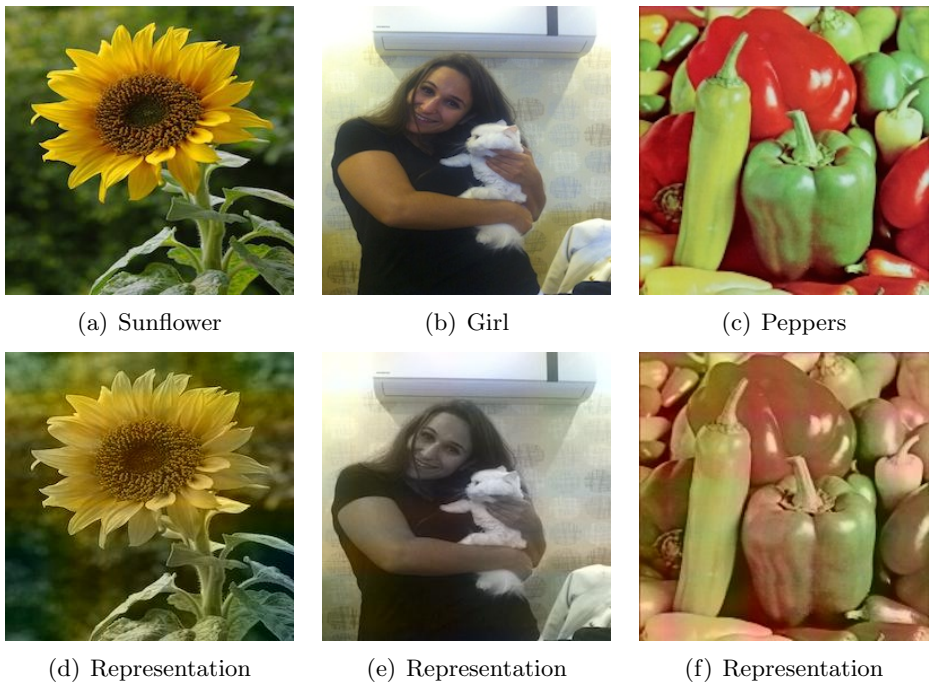
constant, the univariate and the bivariate terms.

In Figure 3. Some colored images and their representations by obtaining with HDMR are given to understand the efficiency of the method. These representation is created by using all constant, univariate and bivariate of HDMR expansion.

5 Conclusion

In this study, we chose some images and we represented these images by using HDMR method. Our aim is to investigate the performance of HDMR method in image reconstruction. Our research shows that if the original image is grayscale then the method works very well. But if we have a colored image in RGB format, then although many features of the image are captured we cannot represent the images exactly. We left to find the way of improving the performance of HDMR for this purpose as a future work. Because if we do that, we can use the HDMR method as an image reconstruction method also for colored images.

We use HDMR here because we decompose three-dimensional array to a constant value, some vectors and some matrices. Then, we create new images by using superpositioning of the constant, vectors and matrices. According to the results, the constant and univariate terms of HDMR are not sufficient to demonstrate the image under consideration. When the bivariate terms are used, then the pattern and colors of the image are determined. To



(a) Sunflower (b) Girl (c) Peppers

(d) Representation (e) Representation (f) Representation

Figure 3: Some original images and their HDMR representations

understand at what level the image obtained through HDMR represents the original one, we evaluate the relative error values.

To get the approximants, relative error values related to these approximations and finally the images, we wrote several codes on MATLAB [12] and all results are performed on a computer of Intel Core i5 1.3 GHz processor and 4GB 1600 MHz DDR3 RAM.

References

[1] I.M. Sobol, Sensitivity Estimates for Nonlinear Mathematical Models. *Mathematical Modelling and Computational Experiments (MMCE),* **1**,(1993) 407–414.

[2] M. Demiralp, High Dimensional Model Representation and its Application varieties. *Mathematical Research,* **9**, (2003) 146–159.

[3] M.A.Tunga and M.Demiralp, *A new approach for data partitioning through high dimensional model representation*, Int. Journal of Computer Mathematics **12** (2008) 1779–1792.

[4] B.N. Rao and R.Chowdhury, *Factorized high dimensional model representation for structural reliability analysis*, Engineering Computations **25** (2008) 708–738.

[5] G. Li, S. Wang, C. Rosenthal and H. Rabitz, *High Dimensional Model Representations Generated from Low Dimensional Data Samples I. mp-Cut-HDMR*, Journal of Mathematical Chemistry **30** (2001) 1–30.

[6] Z. Liao, H. Hoppe, D. Forsyth and Y. Yu , *A Subdivision-Based Representation for Vector Image Editing*, IEEE Transactions on Visualization and Computer Graphics **18** (2012) 1858–1867.

[7] X. Wu, B. Wang, Y. Xia , *A New Method for Image Representation*, Communication Software and Networks (ICCSN) (2011) 138–144.

[8] W. Lihai, Y. Xiangfei and Y.Zaixing, *A Preliminary Study on Two Dimensional Image Reconstruction of Log Cross-section Defect Based on Stress Wave*, International Conference on Measuring Technology and Mechatronics Automation **1** (2010) 291–294.

[9] R. Khilar, DR. S. Chitrakala and S. SelvamParvathy, *3D Image Reconstruction: Techniques, Applications and Challenges*, Proceedings of International Conference on Optical Imaging Sensor and Security (2013) 1–6.

[10] Gonzalez, R. C., Woods, R. E., *Digital Image Processing*, Addison-Wesley: Reading, Massachusetts, 2002.

[11] C. Denker, A. Tritschler and M. Lfdahl, *Image Reconstruction*, Encyclopedia of Optimal Engineering, Marcel Dekker Inc. (2004) 1–19.

[12] MATLAB, *version 7.10.0 (R2010a)*, The MathWorks Inc. , 2010.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

A parallel algorithm for secure multicast

J.A. Alvarez-Bermejo[1], J.M. Arrufat[2] and J.A. Lopez-Ramos[3]

[1] *Department of Informatics, University of Almeria*

[2] *Corporate and investment banking, Management Solutions*

[3] *Department of Mathematics, University of Almeria*

emails: jaberme@ual.es, jose.manuel.arrufat@msspain.com, jlopez@ual.es

Abstract

In this work we introduce a way to parallelize a classical secure multicast protocol that nowadays is unused due to computation and data management requirements when the audience gets large, although it has nice cryptographic properties, showing its actual applicability.

Key words: cryptography, secure multicast, parallel algorithm
MSC 2000: AMS 94A60

1 Introduction

Protection of information that is sent through an unsecure channel is a classical problem, but nowadays is of particular interest due to the massive use of Internet for communications and particularly for information distribution. The increasing interest of distribution of contents in streaming, for instance IPTV, makes necessary to find methods that allow to develop this task in an efficient way.

The so-called secure multicast protocols (cf. [12]) are nowadays the preferred solution. These allow a user to send contents to a plurality of users in a secure and efficient way and every user recovers the original information using a common session key that is renewed every time a user leaves or joins the communication group that is sharing the information. These kind of methods are widely extended in applications such as IPTV ([9]).

The secure multicast protocol introduced in [5] and known as Secure Lock, shows a great simplicity and that its parameters concerning communication overhead are better

than most of those used nowadays. However, computational requirements at the server side are resource demanding and it turns out inefficient as the number of users increases, (cf. [6]), as is the case of most of applications that pretend to be developed by these type of schemes. Efficiency problems concerning computational requirements are due to the use of the Chinese Remainder Algorithm (CRA) to generate rekeying messages. We note that the aforementioned method given in [9] is based on the existence of an interpolator polynomial over a finite field, which can be considered also as an output of a CRA. In that case, inefficiency is solved by distributing users into groups on a tree arrangement. Other cryptographical applications making use of CRA are, for instance those given in [8] and [10].

Thus it is easily understandable the interest in obtaining an optimization of CRA. Our purpose in this paper is to offer an efficient implementation of Secure Lock in order to make it feasible to be used in actual applications and whose ideas could be also applied to other settings.

2 The Secure Lock

Let us recall from [5] the definition of the so-called Secure Lock protocol. So let $\mathcal{U}_i$, $i = 1, \ldots, n$ be a group of users and let k_i and m_i, $i = 1, \ldots, n$ be a key corresponding to any symmetric block cryptosystem and an integer such that m_i and m_j are coprime whenever $i \neq j$. Every user $\mathcal{U}_i$ owns a pair $(k_i, m_i$.

Now let S be the secret that a server aims to distribute among the users $\mathcal{U}_i$, $i = 1, \ldots, n$. To this end it acts as follows:

- Encrypts the secret S with every secret key k_i, computing $s_i = E_{k_i}(S)$, $i = 1, \ldots, n$, where E_{k_i} denotes the encryption (decryption) function using the key k_i.
- Solves the system of congruences $x \equiv_{m_i} s_i$, getting a solution L.
- Broadcasts L.

When every user gets L he just have to compute $L \equiv_{m_i} s_i$ and then, $S = E_{k_i}(s_i)$.

It is clear that security of the precedent is based on the secrecy of the pair (k_i, m_i) for every $i = 1, \ldots, n$. However as it is easily observed and pointed out in [6], serious problems come out as the number n grows, which is usual in many of the possible applications for multicast of this protocol nowadays and thus the only suggested solution is a distribution of users on a tree arrangement in order to decrease the number of congruences to be solved (cf [6]). However it is very common that any broadcasting could reach millions of users, what means that the considered tree should have a considerable depth, since its degree should be small in order not to solve systems with a large number of congruences. This implies that

the number of keys stored at every user's side should be large (one more than the depth of the considered tree).

If we try to parallelize the process of finding solutions to the corresponding system of congruences following what it is made in [10], then the used formula to do that is

$$S = \sum_{i=1}^{n} a_i \cdot (\prod_{i=1}^{n} m_i) \cdot (\prod_{i=1, i \neq j}^{n} m_i)^{-1} (mod\ m_j)$$

and consequently we will face off the problem of dealing with large integers that are given by the products and thus we will may find problems related to storing due lack of memory.

3 An efficient implementation

It is clear that first step of the algorithm where the secret is encrypted using every private key is clearly solved by parallelizing the process. However, as it was previously set, problems arise in the CRA step when trying to get the solution of the corresponding system of congruences. Our first approach could be trying to parallelize this CRA step by using a divide and conquer strategy. However this is not applicable to every architecture or every parallel implementation environment. Thus we need to get an approach that we could use in a generic way. To do so we need to avoid recursions, that require a big amount of computational resources. To do so, we will assume that the set of congruences is already divided in an efficient way. In practice this implies a deduction based on the following: we have as many congruences as the size of the system. Considering this we are able to define combinations of the elements of the set and thus we can define groups inside our set and therefore we are avoiding the process of creating a tree for the resolutions of the system of congruences that would start from the leaves. We propose then the following two algorithms (algorithms 1 and 2):

Algorithm 1:

---

1. Limit = n÷2
2. **for** $i = 0$ to $\log_2$n **do**
3. Resolution (Coefficients, Modules, Limit)
4. Limit = Limit ÷ 2
5. Actualize sets of coefficients and modules

6. **end for**
7. **return** solution

Algorithm 2:

---

1. **for** $i = 0$ to $i = Limit$, run in parallel **do**
2. ResolutionCRA(Coefficients, Modules, 2∗i, 2∗i + 1)
3. **end for**
4. **return** solution

4 Tests

We have made a comparison of three possible implementations. Firstly, the above considered in [10]; secondly, a classical recursive implementation of CRA and finally, the one given by Algorithms 1 and 2. These implementations have been made taking into account two different approaches in parallel computing, that are completely opposite. In the first two cases we have chosen a CPU based parallelization, making use the multicore technology of the processor i7-920, described below (see Table 1), together with the power of the set of resources Parallels .NET. Particularly, in the recursive case we have chosen an hybrid platform CPU-GPGPU, more precisely the above noted processor supported by GPU NVIDIA 260GTX, in order to use the multicore capabilities of the CPU jointly with the power of the GPU platform.

	CPU Intel Core i-7	GPU ASUS EN260GTX
Processors	1	9
Cores/proc	4	24
Frequency	2.66 Ghz	576 Mhz
RAM	4096 MB DDR 1333MHz	896 MB DDR3 2000MHz

Table 1: Characteristics of the execution platform

As we can see in Table 1 the CPU system shows more powerful computation capabilities than the GPU, whereas the latter offers many processing units. With these configurations, we have developed a set of tests over words of length 16 and 32 bits and over sets of 100, 1000 and 10000 congruences, results can be found in Table 2, Table 3 and Table 4.

 ISBN: 978-84-616-9216-3

nBits	GPU*	CPU FB*	CPUyDV
16	3.04	0.08	0.4
32	3.34	0.10	0.64

Table 2: Medium times for 100 executions, measured in ms for a population of 100

As we can observe the algorithm based on that of [10] shows the best results, but with the problems above noted on the memory storing. This is due to the under-utilization of the GPU platform.

nBits	GPU*	CPU FB*	CPUyDV
16	4.20	0.10	16.32
32	4.30	0.15	38.80

Table 3: Medium times for 100 executions, measured in ms for a population of 1000

As the number of users grows we can observe the problems of the algorithm considered in [10] and the hybrid implementation shows a better performing. More precisely, for 32 bits length words we get a speedup of 9x.

nBits	GPU*	CPU FB*	CPUyDV
16	15.54	0.12	19.75
32	16.20	0.2	102.77

Table 4: Medium times for 100 executions, measured in ms for a population of 10000

A progressive increasing of the population clearly slow-down time execution on the CPU platform whereas the hybrid case keeps showing acceptable execution times with a speedup of 7x.

Acknowledgements

First author is supported by grant TIN2008-01117. Third author is supported by grant FQM211 of Junta de Andalucía.

References

[1] C. W. Misner, K. S. Thorne and J. A. Wheeler, *Gravitation*, Freeman, San Francisco, 1970.

[2] E. Witten, *Supersymmetry and Morse theory*, J. Diff. Geom. **17** (1982) 661–692.

[3] N. Antequera and J.A. Lopez-Ramos Remarks and countermeasures on a cryptoanalysis of a secure multicast protocol. *Proceedings of 7th International Conference on Next Generation Web Services Practices, Salamanca 2011*, pp. 201–205, 2011.

[4] H. Chen CRT-Based High-Speed Parallel Architecture for Long BCH Encoding. *Circuits and Systems II: Express Briefs, IEEE Transactions on*, 2009, vol.56, 8, pp. 684–686. issn 1549-7747

[5] G. Chiou and W. Chen Secure broadcasting using the secure lock. *IEEE Trans. Softw. Eng.*, vol. 15(8), pp. 929–934, 1989.

[6] S. Kruus and J. P. Macker. Techniques and issues in multicast security. *Proceedings of Military Communications Conference, MILCOM*, pp. 1028–1032, 1998.

[7] K.-Y. Lin and B. Krishna and H. Krishna Rings, fields, the Chinese remainder theorem and an extension-Part I: Theory. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol.41(10), pp. 641–655. issn 1057-7130, 1994

[8] K.-Y. Lin and B. Krishna and H. Krishna Rings, fields, the Chinese remainder theorem and an extension-Part II: applications to digital signal processing. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 1994,nth=oct, vol.41, 10, pp. 656–668. issn 1057-7130

[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*, vol. 50(2), pp. 632–637, 2004.

[10] Y. Li; Limin Xiao and Z. Wang and H. Tian High Performance Point-Multiplication for Conic Curves Cryptosystem Based on Standard NAF Algorithm and Chinese Remainder Theorem *Information Science and Applications (ICISA), 2011 International Conference on* , pp.1-8, 26-29 April 2011

[11] H. Toyoshima and K. Satoh and K. Ariyama High-speed hardware algorithms for Chinese remainder theorem. *Circuits and Systems, 1996. ISCAS '96., Connecting the World., 1996 IEEE International Symposium on*, 265 -268 vol.2

[12] S. Zhu and S. Jajodia Scalable group key management for secure multicast: A taxonomy and new directions. *Network Security. H. Huang, D. MacCallum and D.-Z. Du (eds.) Springer, United States*, pp. 57–75, 2010.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

A new Runge-Kutta-Nyström pair for the numerical solution of periodic initial value problems

Z.A. Anastassi[1] **and A.A. Kosti**[2]

[1] *Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, P.O. Box 2713, Doha, Qatar*

[2] *Jasmine Residence, Ibn Al Jamouh St., Fereej Bin Mahmoud, Doha, Qatar,*

emails: zackanas@gmail.com, athinakosti@hotmail.com

Abstract

We consider the optimization of the embedded RKN 6(4) pair of Moawwad El-Mikkawy, El-Desouky Rahmo. The new method, which has variable coefficients, is constructed by the nullification of the phase-lag and amplification-error of a method based on the previous based pair. We verify the preservation of the algebraic order and evaluate the principal term of the local truncation error. Furthermore, we perform the periodicity analysis and numerical tests to measure the efficiency of the new method via the integration of several initial value problems.

Key words: Initial value problems, numerical solution, Runge-Kutta-Nyström methods, embedded pairs, phase fitting, amplification fitting
MSC 2000: 65L05 and 65L06

1 Introduction

We investigate the solution of second order differential equations of the form $y''(x) = f(x, y)$, while the initial conditions $y(x_0) = y_0, \quad y'(x_0) = y'_0$ hold. This general problem, especially when it presents an oscillatory behavior, is often met in many areas of astronomy, astrophysics, quantum mechanics etc. For the numerical integration of the above problem we consider the class of Runge-Kutta-Nyström (RKN) methods.

The efficiency of RKN methods lies partially on the use of a variable step size, which can be determined by using an algorithm for automatic step size control, usually attained by utilizing an estimation of the local truncation error of the numerical method. Often,

this is achieved by embedding a second RKN method with the same or lower number of stages and lower order. Dormand, El-Mikkawy and Prince, by setting and satisfying various criteria, have developed highly efficient algorithms, like the 6(4) pair, developed in [2] with six stages that owns the FSAL (First Stage As Last) property, thus effectively using five stages. Despite the fact that the FSAL property lowers the computational cost, it also restricts the further optimization of the efficiency criteria, by involving additional equations. More recently, El-Mikkawy and Rahmo have created a 6(4) pair with six stages without the FSAL property, which is proved to have improved characteristics and be more efficient than the FSAL pair of Dormand, El-Mikkawy and Prince [3].

The above mentioned algorithms have been developed for the efficient solution of the general initial value problem. However, they discard information about an oscillatory/periodic nature of the problem, utilized by various methodologies. Here we apply the methodology of phase-fitting and amplification-fitting to the 6(4) pair of El-Mikkawy and Rahmo [3]. In order to achieve this, the new method must have two variable coefficients that depend on the product of the dominant frequency of the problem and the step-length. In this way, the produced pair with variable coefficients, will succeed zero phase-lag and amplification-error. Additionally, we evaluate the local truncation error for the new method, the corresponding method of El-Mikkawy and Rahmo and the method of Dormand, El-Mikkawy and Prince. Furthermore, the periodicity analysis of the high-order method of the new pair reveals that the latter is "almost" P-stable, in the sense that is P-stable excluding some discrete values. The numerical results prove the efficiency of the new pair via the integration of several oscillatory Initial Value Problems (IVPs).

2 Basic theory

2.1 Explicit Runge-Kutta-Nyström methods

The general form of an explicit $s-$stage Runge-Kutta-Nyström method is presented in (1).

If y_{n+1} and $y_{n+1}^{'}$ denote the approximations of $y(x_{n+1})$ and $y^{'}(x_{n+1})$ respectively, where $x_{n+1} = x_n + h$, $n = 0, 1, \ldots$, then for the numerical solution of the general problem we have the following algorithm

$$\begin{cases} y_{n+1} = y_n + h\, y_n^{'} + h^2 \sum_{i=1}^{s} b_i\, k_i, \\ y_{n+1}^{'} = y_n^{'} + h \sum_{i=1}^{s} b_i^{'}\, k_i, \end{cases} \tag{1}$$

where $k_i = f(x_n + c_i\, h,\ y_n + h\, c_i\, y_n^{'} + h^2 \sum_{j=1}^{i-1} a_{ij}\, k_j)$, $i = 1, \ldots, s$,

 ISBN: 978-84-616-9216-3

An embedded $q(p)$ RKN pair consists of two methods, one (c, A, b, b') of order q and another $(c, A, \hat{b}, \hat{b}')$ of order $p < q$. The high order method produces the solution $(y_{n+1},\, y'_{n+1})$, while the low order method produces the solution $(\hat{y}_{n+1},\, \hat{y}'_{n+1})$, which is only used for the estimation of the local truncation error.

Given an initial step length, every next step length h_{n+1} is determined through the following algorithm

$$h_{n+1} = 0.9\, h_n \left(\frac{TOL}{EST}\right)^{\frac{1}{p+1}},$$

where

$$EST = max\{\|\delta_{n+1}\|_\infty,\, \|\delta'_{n+1}\|_\infty\},$$

$$\delta_{n+1} = \hat{y}_{n+1} - y_{n+1}, \quad \delta'_{n+1} = \hat{y}'_{n+1} - y'_{n+1}.$$

TOL represents the maximum allowed local error. If $EST < TOL$, then the step is accepted, otherwise it is rejected and is repeated with a new step length provided by the algorithm above [3].

The algebraic order of a Runge-Kutta-Nyström method is given by the Definition below

Definition 1 *[7] It is said that a Runge-Kutta-Nyström method has algebraic order p if:*

$$\begin{cases} y_{n+1} - y(x_0 + h) = O\left(h^{p+1}\right) & and \\ y'_{n+1} - y'\left(x_0 + h\right) = O\left(h^{p+1}\right), & n = 1,\, 2, \cdots,\, p. \end{cases} \tag{2}$$

2.2 Analysis of phase-lag, amplification error and stability

The analysis of the phase-lag, amplification error and stability of method (1) is based on the test equation

$$y'' = -\omega^2\, y, \quad \omega \in \mathbb{R} \quad \text{with} \quad y(x_0) = y_0 \quad y'(x_0) = y'_0. \tag{3}$$

After the application of method (1) to the scalar test equation (3), we produce the numerical solution

$$\begin{bmatrix} y_n \\ h\, y'_n \end{bmatrix} = \left[M(v^2)\right]^n \begin{bmatrix} y_0 \\ h\, y'_0 \end{bmatrix}, \quad v = \omega\, h, \quad \text{where} \tag{4}$$

The characteristic equation corresponding to the difference equation (4) is

$$\lambda^2 - tr(M(v^2))\, \lambda + det(M(v^2)) = 0 \tag{5}$$

We have the following theorem:

 ISBN: 978-84-616-9216-3

Theorem 1 [9] *For the Runge-Kutta-Nyström method given in (1), after the application in Eq. (3), we have the following formula for the direct calculation of the phase-lag (or dispersion error)* $\Phi(v)$*:*

$$\Phi(v) = v - \arccos\left(\frac{tr(M(v^2))}{2\sqrt{det(M(v^2))}}\right). \tag{6}$$

If $\Phi(v) = O(v^{q+1})$*, then the method is said to be of phase-lag order* q*.*

Definition 2 [9] *For the Runge-Kutta-Nyström method, presented in (1), the quantity*

$$\alpha(v) = 1 - |\lambda|, \qquad where \quad |\lambda| = \sqrt{det(M(v^2))}$$

is called the amplification error *or the* dissipative error*. If* $\alpha(v) = O(v^{r+1})$ *then the method is said to be of amplification error order* r*.*

Furthermore, we study the stability properties of method (4) when applied to equation (3).

Definition 3 *[8] The stability function* $R(v^2)$ *of the RKN method is defined as the spectral radius* $\rho(M(v^2))$*.*

Definition 4 *[8][1] The interval* $(0, K)$*,* $K \in \mathbb{R}^+ \cup \{+\infty\}$*, so that* $v^2 \in (0, K)$ *is called*

1. *the interval of stability of the RKN method, if* $K = k_{stab}$ *is the highest value such that* $R(v^2) < 1$*,*

2. *the interval of periodicity of the RKN method, if* $K = k_{per}$ *is the highest value such that* $R(v^2) = 1$ *and* $\left[tr\left(M(v^2)\right)\right]^2 - 4\,det\left(M(v^2)\right) < 0$ *(the eigenvalues of* M *are complex conjugate).*

- *If* $(0, k_{stab})$ *lies in the stability interval, then* k_{stab} *is called stability boundary.*
- *If* $(0, k_{per})$ *lies in the periodicity interval, then* k_{per} *is called the periodicity boundary.*
- *If* $k_{stab} = \infty$*, then the RKN method is A-stable [11].*
- *If* $k_{per} = \infty$*, then the RKN method is P-stable [6].*

 ISBN: 978-84-616-9216-3

3 Construction, analysis and application of the new pair

We consider the 6(4) embedded explicit Runge-Kutta-Nyström pair of M. El-Mikkawy, E.D. Rahmo [3], which has 6 stages and does not use the FSAL property. Based on this pair, we develop an optimized pair, where the high order method has zero phase-lag and zero amplification error.

For the development of the optimized method, we set b_2 and b_3 free and then nullify the phase-lag $\Phi(v)$ and the amplification error $\alpha(v)$. These two expressions depend now on b_2 and b_3, apart from v, thus we can nullify them by solving for b_2 and b_3. The solution of the system $\{\Phi(v) = 0, \alpha(v) = 0\}$ yields the new method with $b_2(v)$ and $b_3(v)$. Of course $b_2(0)$ and $b_3(0)$ are identical to the constant coefficients b_2 and b_3 of the corresponding classical method.

The local truncation error analysis reveals that the algebraic order of the new method is six. Furthermore, by evaluating the characteristic roots of the new fitted method, it is proved to be "almost" P-Stable (except for a set of discrete values).

In order to measure the efficiency of the method constructed, we compare it to other well known methods, by integrating several oscillatory initial value problems

References

[1] I. Alonso-Mallo, B. Cano, M.J. Moreta, Stability of Runge-Kutta-Nyström methods, Journal of Computational and Applied Mathematics 189 (2006) 120-131.

[2] J.R. Dormand, M.E.A. El-Mikkawy and P.J. Prince, Families of Runge-Kutta-Nyström formulae, IMA Journal of Numerical Analysis 7 (1987) 235-250.

[3] M. El-Mikkawy, E. Rahmo, A new optimized non-FSAL embedded Runge-Kutta-Nyström algorithm of orders 6 and 4 in six stages, Applied Mathematics and Computation 145 (2003) 33-43.

[4] J.M. Franco, I. Gómez and L. Róndez, Four-stage symplectic and P-stable SDIRKN methods with dispersion of high order, Numerical Algorithms 26 (2001) 347-363.

[5] J.M. Franco, M. Palacios, High-order P-stable multistep methods, Journal of Computational and Applied Mathematics 30 (1990) 1-9.

[6] E. Hairer, Unconditionally stable methods for second-order differential equations, Numerische Mathematik 32 (1979) 373-379.

[7] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations, I, Nonstiff Problems, Springer-Verlag, 2ed, 2008.

 ISBN: 978-84-616-9216-3

[8] B. Paternoster, M. Cafaro, Computation of the interval of stability of Runge-Kutta-Nyström methods, Journal of Symbolic Computation 25 (1998) 383-394.

[9] T.E. Simos, Chemical Modelling - Applications and Theory Vol.1, Specialist Periodical Reports, The Royal Society of Chemistry, Cambridge 1 (2000) 32-140.

[10] E. Stiefel, D.G.Fr. Bettis, Stabilization of Cowell's method, Numerische Mathematik 13 (1969) 154-175.

[11] P.J. Van der Houwen, B.P. Sommeijer, N.H. Cong, Stability of collocation-based Runge-Kutta-Nyström methods, BIT 31 (1991) 469-481.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Evaluation of time-series of satellite reflectance data for land delimitation using clustering algorithms

R. B. Arango[1], I. Díaz[1], A. M. Campos[1], E. F. Combarro[1] and E. R. Canas[2]

[1] *Computer Science Department, University of Oviedo, Spain*

[2] *Technical Direction, Bodegas Terras Gauda, Spain*

emails: rbarango@gmail.com, sirene@uniovi.es, campos@uniovi.es, efernandezca@uniovi.es, ecanas@terrasgauda.com

Abstract

The delimitation of crop land areas grouping zones that share similar soil properties is a key factor in the precision agriculture context. However automatic land delimitation is a challenging task. We propose automatically delimit the zones based on remoted sensed reflectivity and we study how the temporal resolution affects to this delimitation. In order to obtain this zoning, the Partition Around Medoids clustering algorithm has been used and applied to data collected from Terras Gauda vineyar, a well known Spanish producer of Albariño wine. The results are promising in the sense that the clusters obtained are consistent to the current land organization and show that the lower temporal resolution, the more compact the clusters.

Key words: Precision Agriculture, Land Delimitation, Clustering, Satellite Data

1 Introduction

The identification of homogeneous zones of crop land areas is a key factor [1] in the precision agriculture context. These management zones (MZ) [2] address spatial variability of crops grouping areas that share similar soil properties in order to apply specific farming practices to each MZ.

The knowledge of the farmer about the crops and soil could be a starting point in zones determination. However, other approaches provide methods for a systematic MZ

identification such as the classification of apparent soil electrical conductivity [3] or the analysis of yield maps [4].

Sensors and onboard satellite instruments related to environmental and Earth observation, measure electromagnetic radiation emitted and reflected by the observed objects. Based on these radiometric data it is possible to obtain valuable variables and indicators from the agriculture perspective [5] such as moisture and soil temperature, the vegetation index or even the kind of vegetation and its health. These remotely sensed data may be used for the estimation of soil properties and the recognition of spatial patterns [6]. However, non-commercial satellite data products with both high spatial and high temporal resolution are not available.

We propose the identification of homogeneous zones of crops based on remoted sensed reflectivity and we study how the temporal resolution affects to the delimitation of MZ [7], providing decision support to select the satellite data product to identify MZ. We test this method with a case study for the grape vine crops of Terras Gauda, a producer from Galicia (Spain).

This paper addresses the automatic delimitation of the land, selecting algorithms to cluster land points characterised by the satellite reflectance data in order to evaluate how temporal resolution affects the clustering. Section 2 provides an overview of the precision agriculture approach. Section 3 explains the satellite data products used in the clustering process. Section 4 describes the PAM clustering algorithm. Section 5 exposes results of the application of clustering algorithms. Finally, Section 6 shows the conclusions and presents some ideas for future work.

2 Precision Agriculture

Precision Agriculture (PA) or Precision Farming takes advantage of Information and Communications Technology (ICT) in order to provide valuable information and services to farmers. The term "precision" implies that these services can be customized to the needs of an area or specific farm plot and their characteristics, such as the spatial variability of the land and the different features at topographic or geological level. For instance, for the application of fertilizers, knowing the soil nutrients concentration and the in-field variability will allow to choose the right amount of fertilizer [8]. In a similar way, pest control can be more efficient by applying pesticides in a localized manner, where necessary, in contrast to its widespread application on the entire crop [9].

PA uses different measuring techniques such as small-size unmanned air vehicles equipped with spectrometers, luxometers or multi-spectral, modified land vehicles or sensor networks in the field. Using data from wireless sensor networks in greenhouse crops, it is possible to develop disease early-warning systems [10] based on models considering leaf moisture, temperature and time factors. Advances in wireless sensors that use technologies such as

 ISBN: 978-84-616-9216-3

RFID and Very Long Range Identification Tag [11] can help mitigate some drawbacks of mobile sensor networks [12] as their high cost or low autonomy.

However, the deployment of these measurement systems is not essential for PA. The use of satellite imagery and aerial photographs for measuring the radiation emitted and reflected by the fields in several areas of the electromagnetic spectrum allows to observe multiple variables which affect crops [13] as well.

Among the smart services that PA can provide, automatic land delimitation is one of the most challenging tasks. In fact, clustering and automatic delimitation of agroecozones (geographic zones that share similar ecological and environmental features) may be relevant to determine crops potentially more suitable for each zone. Some approaches to automatic land delimitation have been developed. For example, Kumar et al. in [14] apply the k-means clustering algorithm to remote sensing data obtained from the MODIS-based greenness index [15] and also to the seasonal leaf area index [16]. In the same line, Ortega and Santibaez in [17] systematically delimit crop management zones relying on six soil chemical properties related to fertility. Some years ago, Le Ber in [18] built an expert system prototype which is able to recognize different plots, to estimate the production and to classify villages based on particular patterns on the crop arrangement.

3 Digital Data

Valuable data related to precision agriculture such as vegetation indices, land surface temperature or surface reflectance are collected by MODIS (Moderate Resolution Imaging Spectroradiometer) [19]. Around 70 data products are provided by this instrument operated from TERRA and AQUA satellites [16]. Table 1 shows a sample of the MODIS data products related to precision agriculture. These data products are publicly available by HTTP, FTP and at the NASA Land Processes Distributed Active Archive Center [15]. Usually the data will consist of images on JPG format, XML files and data in a hierarchical format (HDF).

In order to generate the data sets for the MZ identification process, daily surface reflectance data products were considered. Specifically MOD09GQ at 250m of spatial resolution. It includes data about surface reflectance for spectral bands 1 and 2 and other variables to measure the quality of the observations and their coverage [20].

Name	Data Product	Res. (m)	Frequency
MYD09GA	Surface Reflectance Bands 1-7	500 m	Daily
MOD09GQ	Surface Reflectance Bands 1-2	250 m	Daily
MOD11A1	Land Surface Temperature and Emissivity	1000 m	Daily
MOD13Q1	Vegetation Indices	250 m	16 days
MOD15A2	Leaf Area Index - FPAR	1000 m	8 days
MOD14A1	Thermal Anomalies and Fire	1000 m	Daily
MOD44B	Vegetation Continuous Fields	250 m	Annual

Table 1: A sample of MODIS data products related to Precision Agriculture

Column	Description
x	Coordinate x of the data point in the UTM 29 CRS
y	Coordinate y of the data point in the UTM 29 CRS
date	Year + Day number of the year in the format YYYYddd
refl_b01	Reflectivity values from MOD09GQ band 1
refl_b02	Reflectivity values from MOD09GQ band 2
num_observations	The number of observations for this measure
QC_250m	A byte of information about the quality of the measure
NDVI	Normalized Difference Vegetation Index
NDVI_scaled	Normalized Difference Vegetation Index. Scaled [0..255] values

Table 2: Columns of the dataset for the MOD09GQ data product. Spatial resolution: 250 m. Temporal resolution: daily. CRS: UTM 29

NDVI indicator is calculated as the relation between the difference of the values of both Red (Red) and NIR (near infrared) channels and its sum [21].

$$NDVI = (NIR - Red)/(NIR + Red)$$

Negative values correspond to water, clouds or snow since their reflectance in the visible spectrum is greater than the corresponding in near infrared, whilst soil and rocks have values near zero. Ranges between 0.1 and 0.6 are indicators of vegetation. Values above 0.6 correspond to dense vegetation canopy.

4 Clustering Algorithm

As it was described in Section 3, the data used in this approach include information about surface reflectance for spectral bands 1 and 2 with a spatial resolution of 250m obtained

 ISBN: 978-84-616-9216-3

from MODIS [20] and the NVDI index. The purpose of this work is to automatically delimit the zones of the vineyard. In order to obtain this zoning, the PAM clustering algorithm [22] has been used. The main characteristics of this algorithm are the following:

- It is a partitioning algorithm. Thus, it breaks the input data up into groups until some stability condition is reached.
- The number of groups is defined in advance.
- PAM stands for Partition Around Medoids. It tries to find a set of objects called medoids that are centrally located in clusters.
- PAM is an algorithm more robust than K-means because it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances.

The bottleneck of clustering algorithms is to properly select the best clustering distribution. In fact, the evaluation of clustering structures is the most difficult task in clustering algorithms. A large number of ways of evaluating the goodness of a clustering algorithm have been proposed in the literature. In this case, since we have no reference to external information, the method selected to validate the clustering was the Silhouette coefficient [23]. It is based on the comparison of cluster tightness and separation. This silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. The average silhouette width provides an evaluation of the clustering validity, and might be used to select an appropriate number of clusters.

Therefore, to select the optimum k according to the Silhouette coefficient, we follow the procedure described below (suppose that the number of points to cluster is n and that K^* is the maximum number of clusters, which is equal to or less than n):

for $j = 1, K^*$ **do**
  for $i = 1, n$ **do**
    $s(i) = \frac{b(i)-a(i)}{max\{(b(i),a(i)\}}$, with $a(i)$ = average dissimilarity between i and all other points of the cluster to which i belongs and $b(i) = min_C d(i, C), \forall\, cluster\, C$
  end for
  $s^j_{avg} = \frac{\sum_{i=1}^{n} s(i)}{n}$
end for
$k = argmax\{s^j_{avg}\}$

Once the number of clusters is computed, the cluster assignment is retrieved, providing the land delimitation.

 ISBN: 978-84-616-9216-3

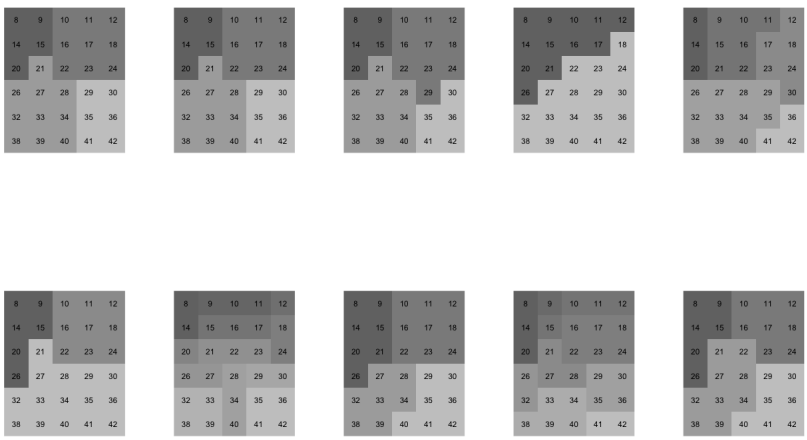


Figure 1: Land delimitation for parcel p_1.

5 Experiments

The purpose of this paper is to automatically delimit the Terras Gauda vineyard. The Terras Gauda vineyard is divided into three separated parcels (hereinafter called p_1, p_2, p_3). According to the spatial resolution provided by the MODIS satellite, that is 250m., p_1 and p_2 are represented using 30 points and p_3 by 16 points. Each point is characterised by 4 variables per day (surface reflectance for spectral bands 1 and 2, NVDI and NVDI-scaled indexes). As data were extracted for 90 days, each point x is represented by 360 values as follows

$$x = (b1_{day_1}, b2_{day_1}, NVDI_{day_1}, b1_{day_2}, b2_{day_2}, NVDI_{day_2}, \ldots, b1_{day_{90}}, b2_{day_{90}}, NVDI_{day_{90}})$$

Following the procedure described in Section 4, we have clustered the three parcels using the Manhattan distance as dissimilarity metric. We have tested the performance of the values obtained from the MODIS satellite in automatic land delimitation and also the effect of different temporal resolutions. To test the temporal resolution we consider all the attributes every day, every 2 days ... until every 10 days. To test the effect of the four attributes, the clustering procedure is performed considering each possible combination of the four attributes obtained per day. However, due to length limitations, only the results obtained when the four attributes afore defined are considered together.

Figures 1 to 2 respectively show the land delimitation for parcels p_1, p_2 an p_3. The structure of each figure is the following: It contains 10 squares divided into smaller squares according to the the spatial resolution provided by MODIS. Considering each figure as a matrix with 2 rows and 5 columns, the square at position $[1, j]$ contains the clusters obtained when temporal resolution is j. The square at position $[2, j]$ contains the clusters obtained when temporal resolution is $j + 5$. Therefore the topleft square represents clustering results

 ISBN: 978-84-616-9216-3

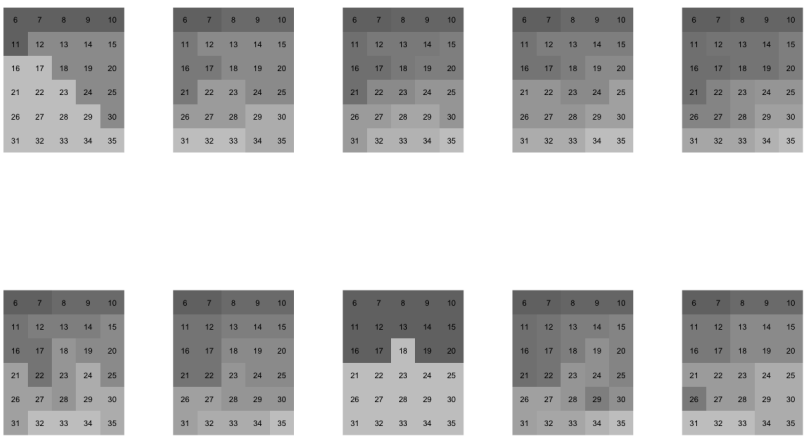


Figure 2: Land delimitation for parcel p_2.

for daily resolution and the bottomright square represents clustering results for a ten days resolution.

The question now is how to decide which land delimitation is the best. As it was stated in Section 4, the criterion used is the Silhouette coefficient. X-axis of Figure 4 represents the temporal resolution (from 1 day to 10 days) while Y-axis represents the value of the Silhouette coefficient. As it can be seen in Figure 4 the lower the temporal resolution, the higher the Silhouette coefficient. Therefore, the main result we can extract is that there is no need to include daily information in order to keep the performance of the clusters. In fact, the clusters are more compact when the temporal resolution is lower.

6 Conclusions and Future Work

This paper presents a first attempt to automatic delimitation of Terras Gauda vineyard. Terras Gauda is a well known Spanish producer of Albariño wine. The results are promising in the sense that the clusters obtained are consistent to the current land organization. In addition, we have checked how the temporal resolution of the data characterizing the land affects the land delimitation. The results show that the lower the resolution, the more compact the clusters. This work opens many interesting new problems. Among them, to study the performance of hierarchical algorithms in automatic land delimitation and to retrieve data from Landsat 8, whose spatial resolution is much higher but the temporal one is lower.

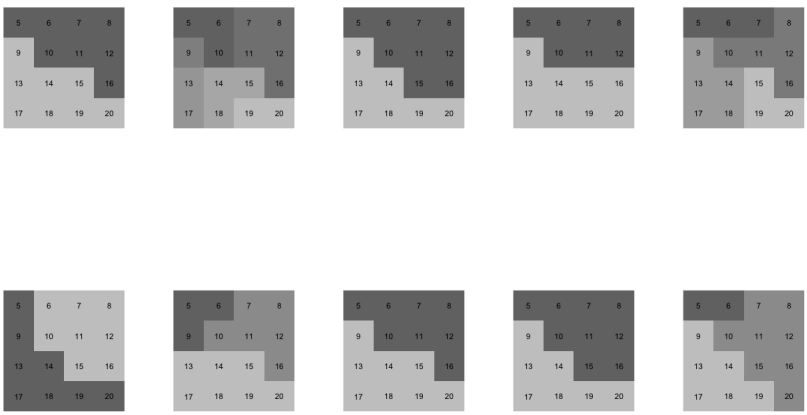


Figure 3: Land delimitation for parcel p_3.

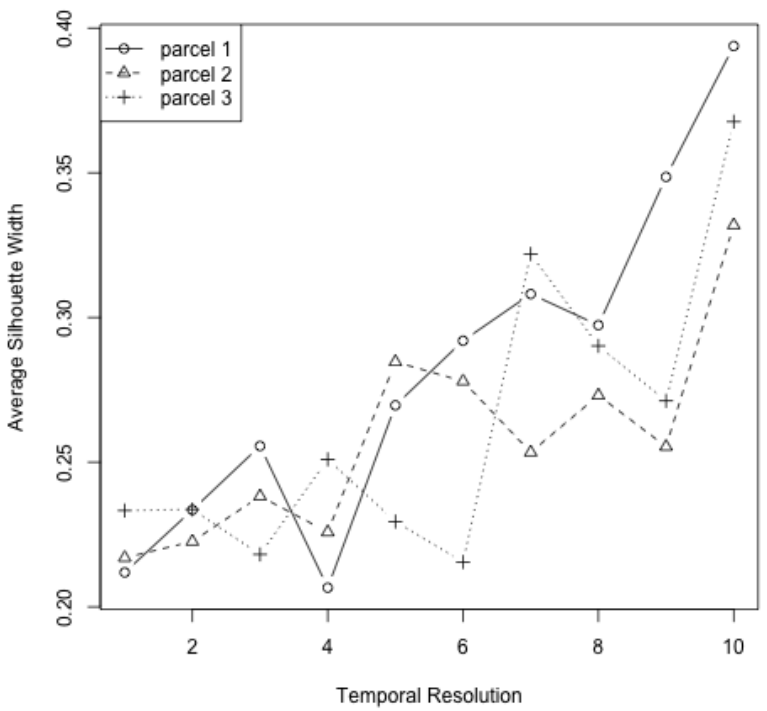


Figure 4: Silhouette coefficient obtained by the clustering algorithm

 ISBN: 978-84-616-9216-3

Acknowledgements

This work has been partially supported under MEC and FEDER Grant TEC2012-38142-C04-04, UNOV-13-EMERG- GIJON-10 grant from University of Oviedo and Pilot B no.621074 from European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under grant agreement no. 621074.

References

[1] A. R. Schepers, J. F. Shanahan, M. A. Liebig, J. S. Schepers, S. H. Johnson, A. Luchiari, Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years, Agronomy Journal 96 (1) (2004) 195–203.

[2] R. Ferguson, R. Lark, G. Slater, Approaches to management zone definition for use of nitrification inhibitors, Soil Science Society of America Journal 67 (3) (2003) 937–947.

[3] C. K. Johnson, D. A. Mortensen, B. J. Wienhold, J. F. Shanahan, J. W. Doran, Site-specific management zones based on soil electrical conductivity in a semiarid cropping system, Agronomy Journal 95 (2) (2003) 303–315.

[4] S. Blackmore, R. J. Godwin, S. Fountas, The analysis of spatial and temporal trends in yield map data over six years, Biosystems engineering 84 (4) (2003) 455–466.

[5] S. Ormeño Villajos, A. Arozarena Villar, M. Martínez Peña, M. Palomo Arroyo, G. Villa Alcázar, J. Peces Morera, L. Pérez García, Los satélites de media y baja resolución espacial como fuente de datos para la obtención de indicadores ambientales, in: IX Congreso Nacional de Medio Ambiente, Madrid, 2008.

[6] A. Bhatti, D. Mulla, B. Frazier, Estimation of soil properties and wheat yields on complex eroded hills using geostatistics and thematic mapper images, Remote Sensing of Environment 37 (3) (1991) 181–191.

[7] F. Zhou, A. Zhang, L. Townley-Smith, A data mining approach for evaluation of optimal time-series of modis data for land cover mapping at a regional level, ISPRS Journal of Photogrammetry and Remote Sensing 84 (2013) 114–129.

[8] H. Yu, D. Liu, G. Chen, B. Wan, S. Wang, B. Yang, A neural network ensemble method for precision fertilization modeling, Mathematical and Computer Modelling 51 (11) (2010) 1375–1382.

[9] C. Fernández-Quintanilla, J. Dorado, C. San Martín, J. Conesa-Muñoz, A. Ribeiro, A five-step approach for planning a robotic site-specific weed management program for

winter wheat, Robotics and Associated High-Technologies and Equipment For Agriculture, 2011.

[10] M. Neto, F. Baptista, L. Navas, G. Ruiz, A business intelligence approach to support a greenhouse tomato crop grey mould disease early warning system, in: T. Mildorf, K. C. jr. (Eds.), ICT for Agriculture, Rural Development and Environment, Czech Centre for Science and Society, 2012, pp. 175–184.

[11] Z. Krivanek, K. Charvat, J. Jezek, M. Musil, Vlit node sensor technology and prefarm, AGRIS on-line Papers in Economics and Informatics 2.

[12] P. Kubíček, V. Lukas, J. Kozel, Selected issues of wireless sensor networks geovisualization in agriculture, in: ICT for Agriculture, Rural Development and Environment, 2012, pp. 249–263.

[13] W. Bingfang, M. Jihua, Z. FeiFei, D. Xin, Z. Miao, C. Xueyang, Applying remote sensing in precision farming-a case study in yucheng, in: World Automation Congress, 2010, pp. 1–6.

[14] J. Kumar, R. T. Mills, F. M. Hoffman, W. W. Hargrove, Parallel k-means clustering for quantitative ecoregion delineation using large data sets, Procedia Computer Science 4 (2011) 1602–1611.

[15] Nasa land processes distributed active archive center (lp daac). aster l1b. usgs/earth resources observation and science (eros) center, sioux falls, south dakota. 2001.

[16] L. DAAC, Modis products table [cited June, 2013].
URL `http://lpdaac.usgs.gov/products/modis_products_table`

[17] R. A. Ortega, O. A. Santibáñez, Determination of management zones in corn (zea mays l.) based on soil fertility, Computers and Electronics in agriculture 58 (1) (2007) 49–59.

[18] F. Le Ber, A prototype model-based expert system for agricultural landscape analysis, AI Applications-Natural Resources, Agriculture, and Environmental Science 9 (2) (1995) 91–101.

[19] L. P. D. A. A. Center, Modis overview [cited October 2013].
URL `https://lpdaac.usgs.gov/products/modis_overview`

[20] Mod09gq [cited May 2014].
URL `https://lpdaac.usgs.gov/products/modis_products_table/mod09gq`

[21] F. Kriegler, W. Malila, R. Nalepka, W. Richardson, Preprocessing transformations and their effects on multispectral recognition, in: Remote Sensing of Environment, VI, Vol. 1, 1969, p. 97.

 ISBN: 978-84-616-9216-3

[22] X. Li, K-means and k-medoids., in: L. Liu, M. T. zsu (Eds.), Encyclopedia of Database Systems, Springer US, 2009, pp. 1588–1589.
URL http://dblp.uni-trier.de/db/reference/db/k.html#Li09f

[23] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (0) (1987) 53 – 65. doi:http://dx.doi.org/10.1016/0377-0427(87)90125-7.
URL http://www.sciencedirect.com/science/article/pii/0377042787901257

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

A hybrid numerical method for a two-dimensional second order hyperbolic equation

Adérito Araújo[1], Cidália Neves[2] and Ercília Sousa[1]

[1] *CMUC, Department of Mathematics, University of Coimbra, Portugal*

[2] *ISCAC, Polytechnic Institute of Coimbra, Portugal*

emails: alma@mat.uc.pt, cneves@iscac.pt, ecs@mat.uc.pt

Abstract

Second order hyperbolic differential equations have been used to model many problems that appear related to heat conduction, mass diffusion and fluid dynamics. In this work a numerical method is presented to solve a two dimensional second order hyperbolic equation with convection terms. A hybrid numerical method is considered which consists of applying the Laplace transform in time and a finite volume discretization in space, where the shape functions associated with the finite volume method are chosen as the combination of hyperbolic functions. We present some numerical tests to show the efficiency of the numerical method.

Key words: Hyperbolic equation, Laplace transform, finite volumes
MSC 2000: AMS codes (35L20, 65M12, 65M22)

1 Introduction

The use of second order hyperbolic differential equations has shown to be useful in modeling diffusive problems. The heat conduction, the mass diffusion and the fluid dynamics are some of the examples belonging to a wide range of subjects covered by these hyperbolic equations. We can find in the literature several proposals for solving these equations for different applications, such as, diffusive problems which include a potential field [4, 6] and various heat conduction problems [2, 5, 9, 10]. However, the incorporation of a convection term in the equation and its effect on the behavior of the solution has not been properly investigated, despite its great relevance in practical applications such as, for instance, the mass concentration distribution of diffusion problems.

We describe briefly the mathematical formulation of the problem under focus. The mass transfer in a two dimensional system is governed by the balance equation

$$\frac{\partial u}{\partial t} + \nabla \cdot J = 0, \tag{1}$$

where u is the mass concentration and J is the mass flux. To accommodate the assumption of finite propagation speed [6, 7, 8] the mass flux verifies the relation

$$J = -\tau \frac{\partial J}{\partial t} - D\nabla u + \mathbf{V}u, \tag{2}$$

where τ is the relaxation time of the mass flux, D the diffusion coefficient and is assumed constant in our study and $\mathbf{V}$ is the velocity field.

Elimination of the mass flux between equation (1) and (2) leads to the hyperbolic equation

$$\tau \frac{\partial^2 u}{\partial t^2} + \frac{\partial u}{\partial t} + \nabla \cdot \mathbf{V}u = D\Delta u. \tag{3}$$

The main purpose of this work is to apply a hybrid numerical method, combining the Laplace transform technique with a finite volume method, to solve the two dimensional second order hyperbolic equation. This hybrid numerical method has been applied in one dimensional problems in [6] and for a pure diffusive problem in two dimensions in [5]. We generalize this numerical method, being an innovation in the context of two dimensional diffusive hyperbolic problems with convection. As we present in this work, the application of the Laplace transform technique is easily generalized to two dimensional problems, but the finite volume discretization requires special attention when we consider partial derivatives of first and second order.

The efficiency of the numerical method is due to the choice of hyperbolic functions used to develop the finite volume method. The method has the advantage of suppressing oscillations, specially when a discontinuity is present in the initial data. Note that for hyperbolic problems the discontinuities may remain through time. Although in [3] an efficient method for one dimensional problems was also introduced to deal with discontinuities, it has the disadvantage of not being generalizable to higher dimensions.

We consider the problem defined by the second order hyperbolic equation

$$\tau \frac{\partial^2 u}{\partial t^2} + \frac{\partial u}{\partial t} + \nabla \cdot \mathbf{V}u = D\Delta u, \quad \mathbf{x} \in \Omega,\ t > 0, \tag{4}$$

where u is the mass concentration, D is the diffusion coefficient, $\mathbf{V}$ is the velocity field and $\tau \in]0, 1]$ is the relaxation time of the mass flux. For our problem we consider the initial conditions given by

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \frac{\partial u}{\partial t}(\mathbf{x}, 0) = u_1(\mathbf{x}), \qquad \mathbf{x} \in \Omega \tag{5}$$

 ISBN: 978-84-616-9216-3

and Dirichlet boundary conditions

$$u(\mathbf{x},t) = f(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega, \ t > 0. \tag{6}$$

Note that for $\tau = 0$, equation (4) is the classical parabolic convection-diffusion equation with initial condition given only by the first equality in (5).

2 The numerical method

The Laplace transform has been used in several works to remove the time dependent terms and obtain a differential equation in space variable ([4], [5], [10], [11]). Using this technique and combining it with an appropriate spatial discretization method has some advantages. First, we can compute the approximate solution for long times accurately and quickly and we do not need to do computations in the time domain using time iterations. Secondly, it also avoids undesirable numerical oscillations that are related with the bad choices of time steps. Any iterative numerical method would take too long to compute the solution for similar times, due to the increased computational effort for discretizing in time, even when we consider an unconditionally implicit numerical method which will allow large time steps. To solve problem (4)–(6) we first apply the Laplace transform to the partial differential equation and boundary conditions, in order to remove the time dependent terms, yielding a differential equation in the space variable that depends on the Laplace parameter. Secondly, we solve the differential equation obtained using a spatial discretization based on a finite volume method that follows an idea presented in [6]. At last, a numerical inverse Laplace transform algorithm is used to obtain the final approximate solution in time and space. The combination of Laplace transform with the finite volume method will be named the Laplace transform finite volume method. We will apply it to our model problem (4)–(6) considering non-trivial initial conditions and different values of the vector $\mathbf{V}$, for both parabolic ($\tau = 0$) and hyperbolic ($\tau \neq 0$) equations.

2.1 One dimensional problem

Let us first see what happens when we consider a one dimensional problem, since it helps to understand the generalization to two dimensions. The problem (4)–(6) can be written as

$$\tau \frac{\partial^2 u}{\partial t^2}(x,t) + \frac{\partial u}{\partial t}(x,t) + \frac{\partial}{\partial x}\left(P(x)u(x,t)\right) = D\frac{\partial^2 u}{\partial x^2}(x,t), \tag{7}$$

where $P(x)$ is now the one dimensional velocity field, with the initial conditions given by

$$u\,(x,0) = u_0(x), \quad \frac{\partial u}{\partial t}\,(x,0) = u_1(x), \qquad x \in (a,b) \tag{8}$$

 ISBN: 978-84-616-9216-3

and the Dirichlet boundary conditions by

$$u\left(a,t\right)=f(t),\qquad u\left(b,t\right)=g(t),\qquad t>0. \tag{9}$$

We denote the Laplace transform of the mass concentration u by $\widetilde{u}$. If we apply the Laplace transform to equation (7) we obtain the ordinary differential equation

$$\frac{d^2\widetilde{u}}{dx^2}(x,s)-\lambda_s^2\widetilde{u}(x,s)-\frac{d}{dx}\left(\frac{P(x)}{D}\widetilde{u}(x,s)\right)=-\frac{u_0(x)}{D}(1+\tau s)-\frac{u_1(x)}{D}, \tag{10}$$

where $\lambda_s=\left(\left(\tau s^2+s\right)/D\right)^{1/2}$ and s is a complex variable, with the boundary conditions, derived from (9), $\widetilde{u}\left(a,s\right)=\widetilde{f}(s)$ and $\widetilde{u}\left(b,s\right)=\widetilde{g}(s)$. The approximate solution of u is obtained by using an inverse Laplace transform algorithm. If P is constant and equation (10) is homogeneous, we are able to apply the inverse Laplace algorithm directly. If we have a non-homogeneous equation, we can apply the inverse Laplace algorithm directly only if we know a particular solution, otherwise we must consider a spatial discretization. If P is non-constant, the spatial discretization is mandatory.

We consider a finite volume formulation to discretize the ordinary differential equation (10). Assume we have a space discretization $x_i=a+i\Delta x,\ i=0,\ldots,N$, where $\Delta x=(b-a)/N$. Let $\widetilde{U}_i(s),\ i=0,\ldots,N$ represent the approximations of $\widetilde{u}\left(x_i,s\right)$ in the Laplace transform domain. After spatial discretization we obtain the linear system

$$K\left(s\right)\widetilde{U}\left(s\right)=\widetilde{b}\left(s\right), \tag{11}$$

where $K(s)=[K_{i,j}(s)]$ is a banded matrix of size $(N-1)\times(N-1)$, with bandwidth three, the vector $\widetilde{U}\left(s\right)$ is given by $\widetilde{U}\left(s\right)=[\widetilde{U}_1\left(s\right),\ldots,\widetilde{U}_{N-1}\left(s\right)]^T$ and $\widetilde{b}\left(s\right)$ contains source terms and boundary conditions.

In what follows, we describe the spatial discretization and give the entries of the matrix K and the vector $\widetilde{b}$. The matrix K and the numerical approximation of the grid point depend on s. However, for the sake of clarity we omit the parameter s denoting $K_{i,j}(s)$ and $\widetilde{U}_i(s)$ by $K_{i,j}$ and $\widetilde{U}_i$ respectively.

The discretization consists of using the finite volume formulation by integrating in x the ordinary differential equation (10) in the i-th control volume $[x_i-\Delta x/2,x_i+\Delta x/2]$,

$$\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}}\left[\frac{d^2\widetilde{u}}{dx^2}-\lambda_s^2\widetilde{u}-\frac{d}{dx}\left(\frac{P}{D}\widetilde{u}\right)\right]dx=-\frac{1}{D}\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}}((1+\tau s)u_0(x)+u_1(x))dx. \tag{12}$$

We compute the integral on the right hand side by the midpoint rule, that is,

$$\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}}((1+\tau s)u_0(x)+u_1(x))dx\simeq\Delta x\left[(1+\tau s)u_0(x_i)+u_1(x_i)\right].$$

 ISBN: 978-84-616-9216-3

We can write the integral on the left hand side as

$$\left[\frac{d}{dx}\widetilde{U}\left(x,s\right)\right]_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}}-\lambda_s^2\left[\int_{x_i-\frac{\Delta x}{2}}^{x_i}\widetilde{U}\left(x,s\right)dx+\int_{x_i}^{x_i+\frac{\Delta x}{2}}\widetilde{U}\left(x,s\right)dx\right]$$
$$-\frac{P\left(x_i+\Delta x/2\right)}{D}\widetilde{U}\left(x_i+\Delta x/2,s\right)+\frac{P\left(x_i-\Delta x/2\right)}{D}\widetilde{U}\left(x_i-\Delta x/2,s\right). \quad (13)$$

For $x\in[x_i,x_{i+1}]$, $i=0,\ldots,N-1$, we approximate $\widetilde{U}\left(x,s\right)$ by the following combination of hyperbolic functions,

$$\widetilde{U}\left(x,s\right)=\frac{\sinh\left(\lambda_s\left(x_{i+1}-x\right)\right)}{\sinh\left(\lambda_s\Delta x\right)}\widetilde{U}_i(s)+\frac{\sinh\left(\lambda_s\left(x-x_i\right)\right)}{\sinh\left(\lambda_s\Delta x\right)}\widetilde{U}_{i+1}(s),$$

where $\widetilde{U}_i(s)$, $i=0,\ldots,N$, represent the approximations of $\widetilde{u}(x_i,s)$ in the Laplace transform domain. These shape hyperbolic functions have been suggested in [6]. Substituting this approximation in (13) yields

$$\frac{\lambda_s}{\sinh(\lambda_s\Delta x)}\left[\widetilde{U}_{i-1}(s)-2\cosh(\lambda_s\Delta x)\widetilde{U}_i(s)+\widetilde{U}_{i+1}(s)\right]$$
$$-\frac{P\left(x_i+\Delta x/2\right)}{D}\frac{\sinh(\lambda_s\Delta x/2)}{\sinh(\lambda_s\Delta x)}\left(\widetilde{U}_i(s)+\widetilde{U}_{i+1}(s)\right)$$
$$+\frac{P\left(x_i-\Delta x/2\right)}{D}\frac{\sinh(\lambda_s\Delta x/2)}{\sinh(\lambda_s\Delta x)}\left(\widetilde{U}_{i-1}(s)+\widetilde{U}_i(s)\right).$$

Finally, the evaluation of (12) produces the following discretized equations, for $i=1,\ldots,N-1$,

$$K_{i,i-1}(s)\widetilde{U}_{i-1}(s)+K_{i,i}(s)\widetilde{U}_i(s)+K_{i,i+1}(s)\widetilde{U}_{i+1}(s)=-\frac{\sinh(\lambda_s\Delta x)}{D\lambda_s}\Delta x\left[(1+\tau s)u_0(x_i)+u_1(x_i)\right] \quad (14)$$

for

$$K_{i,i-1}(s)=1+P_{i-1/2}\frac{\sinh\left(\lambda_s\Delta x/2\right)}{D\lambda_s},\qquad K_{i,i+1}(s)=1-P_{i+1/2}\frac{\sinh\left(\lambda_s\Delta x/2\right)}{D\lambda_s},$$
$$K_{i,i}(s)=-2\cosh\left(\lambda_s\Delta x\right)-\left(P_{i+1/2}-P_{i-1/2}\right)\frac{\sinh\left(\lambda_s\Delta x/2\right)}{D\lambda_s}, \quad (15)$$

where $P_{i\pm1/2}=P\left(x_i\pm\Delta x/2\right)$. The vector that contains boundary terms is given by

$$\widetilde{b}\left(s\right)=-\frac{\Delta x\sinh(\lambda_s\Delta x)}{D\lambda_s}\begin{bmatrix}(1+\tau s)u_0(x_1)+u_1(x_1)\\(1+\tau s)u_0(x_2)+u_1(x_2)\\\vdots\\(1+\tau s)u_0(x_{N-2})+u_1(x_{N-2})\\(1+\tau s)u_0(x_{N-1})+u_1(x_{N-1})\end{bmatrix}-\begin{bmatrix}K_{1,0}(s)\widetilde{U}_0(s)\\0\\\vdots\\0\\K_{N-1,N}(s)\widetilde{U}_N(s)\end{bmatrix},$$

 ISBN: 978-84-616-9216-3

Thus, equation (14) can be written in the matrix form (11) where the matrix K and the vector $\widetilde{b}$ are defined by the entries given above.

The next step is to determine an approximate solution $U(x_i,t)$ from $\widetilde{U}(x_i,s)$ by using the Laplace inversion numerical method described in [1, 4]. The errors that come from the numerical inversion of Laplace transform are described in [4]. We can prove the spatial discretization error, using the finite volume method, is at least of second order and similarly to what was done in [4] we obtain a second order error convergence for the full numerical method.

2.2 Two dimensional problem

The numerical method described in one dimension is extended in this section to solve the two dimensional hyperbolic diffusion equation, defined in a rectangular domain $\Omega \subset \mathbb{R}^2$,

$$\begin{aligned} & \tau\frac{\partial^2 u}{\partial t^2}(x,y,t)+\frac{\partial u}{\partial t}(x,y,t)+\frac{\partial}{\partial x}(P(x)u(x,y,t))+\frac{\partial}{\partial y}(Q(y)u(x,y,t)) \\ = \quad & D\left(\frac{\partial^2 u}{\partial x^2}(x,y,t)+\frac{\partial^2 u}{\partial y^2}(x,y,t)\right), \qquad (x,y)\in\Omega,\ t>0, \end{aligned} \tag{16}$$

where the velocity field $\mathbf{V}$ is now given by $(P(x),Q(y))$. The initial conditions are given by

$$u(x,y,0)=u_0(x,y), \qquad \text{and} \qquad \frac{\partial u}{\partial t}(x,y,0)=u_1(x,y), \qquad (x,y)\in\Omega, \tag{17}$$

and the Dirichlet boundary conditions are given by

$$u(x,y,t)=f(x,y,t), \qquad (x,y)\in\partial\Omega,\ t>0. \tag{18}$$

Similarly to what has been done in one dimension, we apply the Laplace transform to remove the time dependent terms and obtain the equation

$$\frac{\partial^2\widetilde{u}}{\partial x^2}+\frac{\partial^2\widetilde{u}}{\partial y^2}-\lambda_s^2\widetilde{u}-\frac{\partial}{\partial x}\left(\frac{P}{D}\widetilde{u}\right)-\frac{\partial}{\partial y}\left(\frac{Q}{D}\widetilde{u}\right)=-\frac{u_0(x,y)}{D}(1+\tau s)-\frac{u_1(x,y)}{D}, \tag{19}$$

with $\widetilde{u}(x,y,s)$ the Laplace transform of $u(x,y,t)$ and $\lambda_s^2=(\tau s^2+s)/D$. We now generalize the Laplace transform finite volume method presented in the previous section to two dimensions. Consider the control volume $\Omega_{i,j}=[x_i-\Delta x/2,x_i+\Delta x/2]\times[y_j-\Delta y/2,y_j+\Delta y/2]$, $i=1,\ldots,N_x-1$, $j=1,\ldots,N_y-1$, represented in Figure 1 and where the point O represents (x_i,y_j).

We integrate the differential equation (19) within the control volume $\Omega_{i,j}$, that is,

$$\begin{aligned} & \int_{\Omega_{i,j}}\frac{\partial^2\widetilde{u}}{\partial x^2}+\frac{\partial^2\widetilde{u}}{\partial y^2}-\lambda_s^2\widetilde{u}-\frac{\partial}{\partial x}\left(\frac{P}{D}\widetilde{u}\right)-\frac{\partial}{\partial y}\left(\frac{Q}{D}\widetilde{u}\right)dxdy \\ & \quad =-\frac{1}{D}\int_{\Omega_{i,j}}(1+\tau s)u_0(x,y)+u_1(x,y)dxdy. \end{aligned} \tag{20}$$

 ISBN: 978-84-616-9216-3

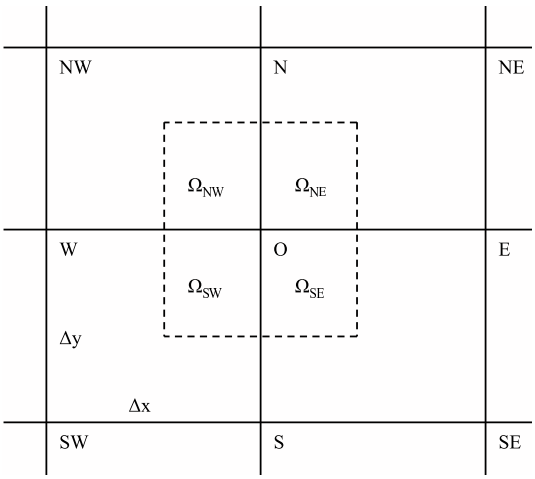


Figure 1: Control volume $\Omega_{i,j}$.

The control volume $\Omega_{i,j}$ is subdivided in four rectangular elements as shown in Figure 1. To derive the discretization, we approximate $\widetilde{u}(x,y,s)$ in terms of the nodal points and the shape functions in each element. The four shape functions are chosen in a similar way to what was done for the one dimensional case, as explained in [5]. For the element Ω_{NE}, and assuming O represents the point (x_i, y_j), the shape functions are given by

$$\begin{aligned}
N_O(x,y,s) &= \frac{1}{\sinh(\mu\Delta x)\sinh(\mu\Delta y)}\sinh(\mu(x_{i+1}-x))\sinh(\mu(y_{j+1}-y)),\\
N_E(x,y,s) &= \frac{1}{\sinh(\mu\Delta x)\sinh(\mu\Delta y)}\sinh(\mu(x-x_i))\sinh(\mu(y_{j+1}-y)),\\
N_N(x,y,s) &= \frac{1}{\sinh(\mu\Delta x)\sinh(\mu\Delta y)}\sinh(\mu(x_{i+1}-x))\sinh(\mu(y-y_j)),\\
N_{NE}(x,y,s) &= \frac{1}{\sinh(\mu\Delta x)\sinh(\mu\Delta y)}\sinh(\mu(x-x_i))\sinh(\mu(y-y_j)),
\end{aligned}$$

where $\mu = \lambda_s/\sqrt{2}$. For this element the solution is then approximated by

$$\begin{aligned}
\widetilde{U}(x,y,s) &= N_O(x,y,s)\widetilde{U}_{i,j} + N_E(x,y,s)\widetilde{U}_{i+1,j} + N_N(x,y,s)\widetilde{U}_{i,j+1}\\
&\quad + N_{NE}(x,y,s)\widetilde{U}_{i+1,j+1}.
\end{aligned}$$

For the other three elements $\widetilde{U}(x,y,s)$ can be represented in a similar way. We compute the integral on the right hand side of equation (20) by the midpoint rule and obtain

$$\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}}\int_{y_j-\frac{\Delta y}{2}}^{y_j+\frac{\Delta y}{2}}((1+\tau s)u_0(x,y)+u_1(x,y))dxdy \simeq \Delta x\Delta y\left[(1+\tau s)u_0(x_i,y_j)+u_1(x_i,y_j)\right].$$

 ISBN: 978-84-616-9216-3

After integration of the left member of (20), the complete discretized equation that corresponds to node O is obtained by the contribution of all the four elements and it originates a compact discretization given by

$$\begin{aligned}
&K_O\widetilde{U}_{i,j} + K_E\widetilde{U}_{i+1,j} + K_W\widetilde{U}_{i-1,j} + K_N\widetilde{U}_{i,j+1} + K_S\widetilde{U}_{i,j-1} + K_{NE}\widetilde{U}_{i+1,j+1} \\
&\quad + K_{NW}\widetilde{U}_{i-1,j+1} + K_{SE}\widetilde{U}_{i+1,j-1} + K_{SW}\widetilde{U}_{i-1,j-1} \\
&= -\frac{\Delta x\Delta y}{D}\sinh(\mu\Delta x)\sinh(\mu\Delta y)\left((1+\tau s)u_0(x_i,y_j) + u_1(x_i,y_j)\right),
\end{aligned}$$

where the coefficients are defined by

$$\begin{aligned}
K_O &= 4[\cosh(\mu\Delta x)\cosh(\mu\Delta y/2) + \cosh(\mu\Delta y)\cosh(\mu\Delta x/2)] - 8\cosh(\mu\Delta x)\cosh(\mu\Delta y) \\
&\quad + \frac{2}{\mu}(P_{i+1/2} - P_{i-1/2})\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y) - \cosh(\mu\Delta y/2)) \\
&\quad + \frac{2}{\mu}(Q_{j+1/2} - Q_{j-1/2})\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x) - \cosh(\mu\Delta x/2)),
\end{aligned}$$

$$\begin{aligned}
K_E &= 2[2\cosh(\mu\Delta y) - \cosh(\mu\Delta y/2) - \cosh(\mu\Delta x/2)\cosh(\mu\Delta y)] \\
&\quad + \frac{2}{\mu}P_{i+1/2}\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y) - \cosh(\mu\Delta y/2)) \\
&\quad + \frac{1}{\mu}(Q_{j+1/2} - Q_{j-1/2})\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x/2) - 1),
\end{aligned}$$

$$\begin{aligned}
K_W &= 2[2\cosh(\mu\Delta y) - \cosh(\mu\Delta y/2) - \cosh(\mu\Delta x/2)\cosh(\mu\Delta y)] \\
&\quad - \frac{2}{\mu}P_{i-1/2}\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y) - \cosh(\mu\Delta y/2)) \\
&\quad + \frac{1}{\mu}(Q_{j+1/2} - Q_{j-1/2})\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x/2) - 1),
\end{aligned}$$

$$\begin{aligned}
K_N &= 2[2\cosh(\mu\Delta x) - \cosh(\mu\Delta x/2) - \cosh(\mu\Delta x)\cosh(\mu\Delta y/2)] \\
&\quad + \frac{1}{\mu}(P_{i+1/2} - P_{i-1/2})\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y/2) - 1) \\
&\quad + \frac{2}{\mu}Q_{j+1/2}\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x) - \cosh(\mu\Delta x/2)),
\end{aligned}$$

$$\begin{aligned}
K_S &= 2[2\cosh(\mu\Delta x) - \cosh(\mu\Delta x/2) - \cosh(\mu\Delta x)\cosh(\mu\Delta y/2)] \\
&\quad + \frac{1}{\mu}(P_{i+1/2} - P_{i-1/2})\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y/2) - 1) \\
&\quad - \frac{2}{\mu}Q_{j-1/2}\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x) - \cosh(\mu\Delta x/2)),
\end{aligned}$$

 ISBN: 978-84-616-9216-3

$$
\begin{aligned}
K_{NE} &= [\cosh(\mu\Delta x/2)+\cosh(\mu\Delta y/2)-2]+\frac{1}{\mu}P_{i+1/2}\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y/2)-1)\\
&\quad +\frac{1}{\mu}Q_{j+1/2}\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x/2)-1),\\
K_{NW} &= [\cosh(\mu\Delta x/2)+\cosh(\mu\Delta y/2)-2]-\frac{1}{\mu}P_{i-1/2}\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y/2)-1)\\
&\quad +\frac{1}{\mu}Q_{j+1/2}\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x/2)-1),\\
K_{SE} &= [\cosh(\mu\Delta x/2)+\cosh(\mu\Delta y/2)-2]+\frac{1}{\mu}P_{i+1/2}\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y/2)-1)\\
&\quad -\frac{1}{\mu}Q_{j-1/2}\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x/2)-1),\\
K_{SW} &= [\cosh(\mu\Delta x/2)+\cosh(\mu\Delta y/2)-2]-\frac{1}{\mu}P_{i-1/2}\sinh(\mu\Delta x/2)(\cosh(\mu\Delta y/2)-1)\\
&\quad -\frac{1}{\mu}Q_{j-1/2}\sinh(\mu\Delta y/2)(\cosh(\mu\Delta x/2)-1).
\end{aligned}
$$

The matricial formulation of the problem is also given by $K(s)\widetilde{U}(s)=\widetilde{b}(s)$, where the matrix K is now a block matrix and each block is a banded matrix with bandwidth three.

This finite volume difference scheme has accuracy of second order in space as will be confirmed by the numerical results.

3 Numerical tests

In this section numerical results are presented for the two dimensional problem to show the second order convergence rate of the numerical method developed, and called Laplace transform finite volume method (Laplace-FV-2D), and also to illustrate the behavior of the solutions. In order to compare the numerical solution $U_{i,j}(t)=U_{i,j}$, $i=1,\ldots,N_x-1$, $j=1,\ldots,N_y-1$ with the respective exact solution $u(x_i,y_j,t)=u_{i,j}$, we consider two problems.

Problem 1: Consider the problem (16)–(18) for $\tau=1$, $P(x)=Q(y)=0$, defined in $\Omega=(0,\sqrt{8}\pi)\times(0,\sqrt{8}\pi)$, with initial conditions given by $u_0(x,y)=\sin(x/\sqrt{8})\sin(y/\sqrt{8})$, $u_1(x,y)=-(1/2)u_0(x,y)$ and boundary conditions $u(x,y,t)=0$ for $(x,y)\in\partial\Omega$, $t>0$. The exact solution is given by $u(x,y,t)=\mathrm{e}^{-t/2}\sin\left(x/\sqrt{8}\right)\sin\left(y/\sqrt{8}\right)$.

Problem 2: Consider the problem (16)–(18), for $\tau=0$, $P(x)=Q(y)=1$, defined in $\mathbb{R}^2$ with initial condition $u_0(x,y)=\mathrm{e}^{-\left(x^2+y^2\right)}$ and assuming $u(x,y,t)=0$ for any (x,y,t) with large (x,y). The exact solution is given by $u\,(x,y,t)=(1/\sqrt{1+4t})\mathrm{e}^{-((x-Pt)^2+(y-Qt)^2)/(1+4t)}$.

To have information about the rate of convergence of the numerical method, we present

 ISBN: 978-84-616-9216-3

$\Delta x = \Delta y$	Problem 1	Rate		$\Delta x = \Delta y$	Problem 2	Rate
$\sqrt{8}\pi/40$	0.3700×10^{-2}			$20/40$	0.1100×10^{-2}	
$\sqrt{8}\pi/80$	0.9349×10^{-3}	2.0		$20/80$	0.2701×10^{-3}	2.0
$\sqrt{8}\pi/120$	0.4156×10^{-3}	2.0		$20/120$	0.1198×10^{-3}	2.0
$\sqrt{8}\pi/160$	0.2338×10^{-3}	2.0		$20/160$	0.6734×10^{-4}	2.0
$\sqrt{8}\pi/200$	0.1496×10^{-3}	2.0		$20/200$	0.4307×10^{-4}	2.0

Table 1: Errors and rates obtained for $t = 1$, $TOL = 1/N^3$, $T = 20$, $\beta = -\ln(10^{-16})/2T$, computed with the norm ℓ_∞. Problem 1: $0 \leq x, y \leq \sqrt{8}\pi$. Problem 2: $-10 \leq x, y \leq 10$.

in Table 1, the ℓ_∞ error norm, defined by

$$\|u - U\|_\infty = \max_{1 \leq i \leq N_x - 1, 1 \leq j \leq N_y - 1} |u(x_i, y_j, t) - U(x_i, y_j, t)|.$$

The results show a convergence rate of second order for Problem 1 and Problem 2.

To illustrate the behaviour of the solutions, we consider two additional problems. Both problems are for $\tau = 1$ and different values of P and Q.

Problem 3: We first consider the problem defined in the domain $\Omega = (0, 1) \times (0, 1)$, with the initial conditions $u_0(x, y) = u_1(x, y) = 0$ and boundary conditions given by $u(x, 0, t) = 0$, $u(x, 1, t) = 0$, $u(0, y, t) = \sin(\pi y)$, $u(1, y, t) = 0$. In Figure 2 we compare the performance of the method we are presenting, the Laplace-FV-2D, with the Laplace transform finite differences method (Laplace-FD-2D). This method is presented in [4] for the one dimensional case and can be easily extended to two dimensions. We observe the Laplace-FV-2D method suppresses oscillations easier than the Laplace-FD-2D method.

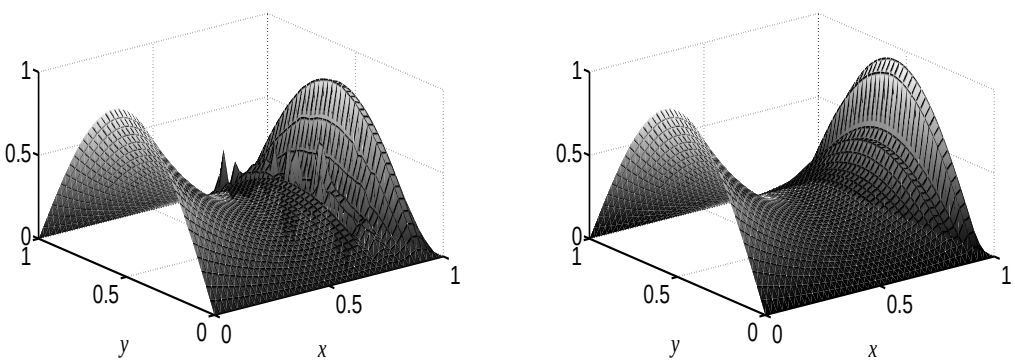


Figure 2: Approximate solution of Problem 3 for $\tau = 1$, $P(x) = 1$ and $Q(y) = 0$ at $t = 1$. Computed with $\Delta x = \Delta y = 0.025$. Left: Laplace-FD-2D. Right: Laplace-FV-2D.

 ISBN: 978-84-616-9216-3

Problem 4: To see how the Laplace-FV-2D method handles a discontinuity at the initial data, we consider the problem defined in the domain $\Omega = (0,4)\times(0,4)$, with the initial conditions $u_0(x,y) = u_1(x,y) = 0$ and boundary conditions $u(x,0,t) = 0$, $u(x,4,t) = 0$, $u(0,y,t) = 1$, $u(4,y,t) = 0$. This is illustrated in Figure 3. Although the solution presents a jump discontinuity in the initial time, the Laplace-FV-2D method performs quite well without oscillations. The behavior of the solution can be observed as we travel in time.

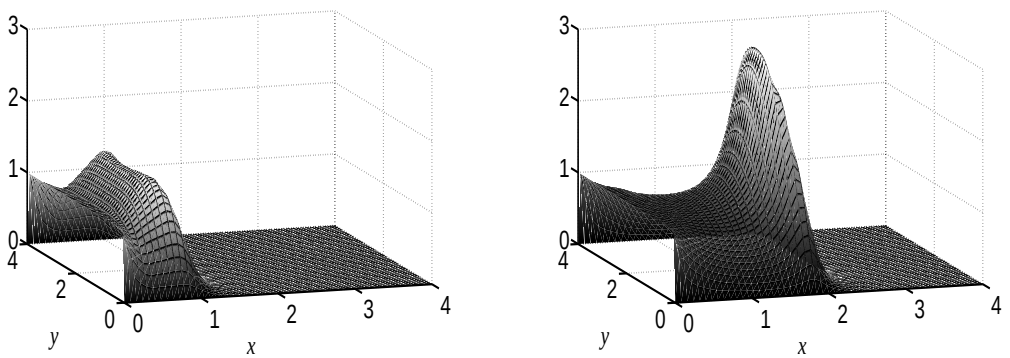


Figure 3: Approximate solution of Problem 4 for $\tau = 1$, $P(x) = 2$ and $Q(y) = 0$. Computed with $\Delta x = \Delta y = 0.08$. Left: $t = 1$. Right: $t = 2$.

4 Final Remarks

We have derived a numerical method to solve a two dimensional hyperbolic problem based on the Laplace transform and the finite volume method. The full technique can be described in three steps. First, we apply the Laplace transform to the partial differential equation and boundary conditions, in order to remove the time dependent terms, yielding a differential equation in the space variable that depends on the Laplace parameter. Secondly, we solve the differential equation obtained using a finite volume method. In the end, a numerical inverse Laplace transform algorithm is used to obtain the final approximate solution in time and space. It has been shown by the numerical results that this numerical method has accuracy of second order, can avoid oscillations and it also deals efficiently with discontinuities that in the case of hyperbolic problems can be propagated through time.

Acknowledgments

This work has been partially supported by CMUC and FCT (Portugal), through European program COMPETE/FEDER.

References

[1] J. Ahn, S. Kang and Y. H. Kwon, *A flexible inverse Laplace transform algorithm and its application*, Computing **71(2)** (2003) 115–131.

[2] Y. M. Ali and L. C. Zhang, *Relativistic moving heat source*, International Journal of Heat and Mass Transfer, **48** (2005) 2741–2758.

[3] A. Araújo, C. Neves and E. Sousa, *A Laplace transform piecewise linearised method for a second order hyperbolic equation*, AIP Conference Proceedings, **1479** (2012) 2187–2190.

[4] A. Araújo, A. K. Das, C. Neves and E. Sousa, *Numerical solution for a non-Fickian diffusion in a periodic potential*, Communications in Computational Physics, **13(2)** (2013) 502–525.

[5] H.-T. Chen and J.-Y. Lin, *Analysis of two-dimensional hyperbolic heat conduction problems*, International Journal of Heat and Mass Transfer, **37(1)** (1994) 153–164.

[6] H.-T. Chen and K.-C. Liu, *Numerical analysis of a non-Fickian diffusion problems in a potential field*, Numerical Heat Transfer, Part B, **40** (2001) 265–282.

[7] A. K. Das, *A non-fiction diffusion equation*, Journal of Applied Physics, **70** (1991) 1355–1358.

[8] R. Kubo, M. Toda and N. Hashitsume, *Statistical Physics II, Nonequilibrium Statistical Mechanics*, Springer, 1995.

[9] H.-C. Lin, M.-I. Char and W.-J. Chang, *Soret effects on non-Fourier heat and non-Fickian mass diffusion transfer in a slab*, Numerical Heat and Mass Transfer, Part A, **55** (2009) 1096–1115.

[10] K.-C. Liu and H.-T. Chen, *Analysis for the dual-phase-lag bio-heat transfer during magnetic hyperthermia treatment*, International Journal of Heat and Mass Transfer, **52** (2009) 1185–1192.

[11] A. Saleh and M. Al-Nimir, *Variational formulation of hyperbolic heat conduction problems applying Laplace transform technique*, International Communications in Heat and Mass Transfer, **35** (2008) 204–214.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Fitted parametrized spline curves help Mapmakers to define roads

F. J. Ariza-López[1], D. Barrera[2] and J. F. Reinoso[3]

[1] *Dpt. of Cartographic, Geodesic and Photogrammetry Engineering, University of Jaén, 23071-Jaén, Spain*

[2] *Department of Applied Mathematics, University of Granada, 18071-Granada, Spain*

[3] *Dpt. of Architectonic and Engineering Graphic Expression, University of Granada, 18071-Granada, Spain*

emails: fjariza@ujaen.es, dbarrera@ugr.es, jreinoso@ugr.es

Abstract

Linear elements are the major object group represented in maps, and its definition is usually expressed as a polygonal shape. Roads are the most important manmade linear elements appearing in cartography, and one of the most important issues which Mapmakers focus in, particularly its positional accuracy. Nowadays there are a lot of cartography sources (National and Regional Agencies, digital cartographic companies, open source organizations, etc), nevertheless the spatial position for the same road can be slowly, even highly, discrepant from each one of the above mentioned source. One could expect getting a better accuracy blending the linear data coming from different sources, i.e., using all the points defining the same road from each cartographic producer. In our study we used a B-spline least square fit in order to improve the positional accuracy for the final result.

Key words: mapmarkers, B-spline, least square fit

1 Introduction

There are several approaches to obtain the trace for the roads: photogrammetry, which is the usual way in medium scale maps, like the MTN25 in Spain (Instituto Geográfico Nacional, 2000), precise differential GPS (Edelkamp and Schrödl, 2003), repeated large traces set coming from different sources or users (Li et al., 2012)... The last approach,

i.e. computing a representative axis from a large set of samples is denoted mining data processing, and particularly in our case is denoted as mining spatial data processing (Lima and Ferreira, 2009). Multiple GPS traces is the preferred dataset to be processed in the axis road determination (Biagioni and Eriksson, 2012,). The methods proposed to estimate the representative axis come from the cluster approach (Edelkamp and Schrdl, 2003) up to the multiples hypothesis for fusioning lines (Schuessler and Axhausen, 2009), passing through cutting the traces set transversely and computing the centroid (Zhang et al., 2010). Any method uses an approach fitting a B-spline to the whole points cloud composed by the traces set (multiple traces captured by a low precise GPS with spatial positional errors around 5 m). We think the main raison because researchers didnt used a B-spline fit is the difficult to program an algorithm, which admits a big points cloud dataset, which, a priori, are unordered.

2 Material, methodology and results

Our data set is composed by 140 road traces defined each one by a polyline. Each trace was captured by a user GPS with spatial accuracy around 5 m. The kind of road we studied can be considered as a medium mountain one. As we can see in figure 1 the road is highly sinuous with some pieces of curves characterized by high curvature. The step we address in order to get a suitable adjustment was as follow:

- Selecting a piece of the same area from each one of the 140 road traces.
- Dividing each piece in the same number of segments, each one by 5 m approximately, so that the resulting points on one line can be seen as the homotopy transformation from whatever of the other lines.
- Grouping the homotopic points in differentiated subsets.
- Fitting a 3D curve in the space of cubic splines by least square method taking into account that each point in a differentiated subset has the same evaluation value. In this linear space we used the B-spline basis associated with the extended partition with multiple knots of a uniform partition of the domain of the parametrized curve (de Boor, 2001).

After fitting the parametrized curve expressed in terms of B-spline, the estimated road axis looks pretty representative for the whole traces set (see figure 1). Previously we had experienced with other methods based on mean values which, although they were suitable, they also produce a few undesirable corners. With the method exposed in this study that problem doesnt appear.

 ISBN: 978-84-616-9216-3

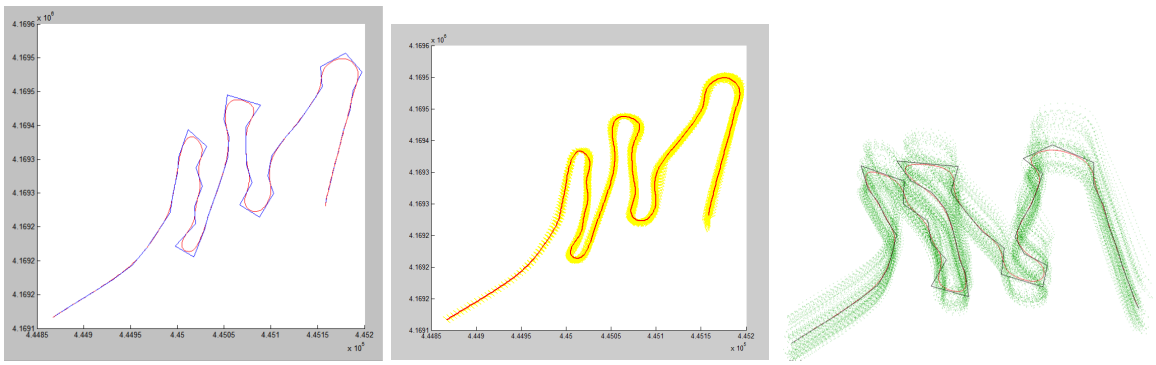


Figure 1: Fitted spline curve resulting: left image contains the B-spline in red and the control points are in blue; central image is a horizontal view including the points cloud in yellow; the right image is a perspective view with the points cloud in green

References

[1] J. Biagioni and J. Eriksson, *Inferring road maps from GPS traces: survey and comparative evaluation*, Transportation Research Record: Journal of the Transportation Research Board **2291** (2012) 61–71.

[2] S. Edelkamp and S. Schrödl, *Route planning and map inference with Global Position Traces*, Comp. Sci. In Perspective, LNCS 2598, pages 128–151, 2003.

[3] Instituto Geográfico Nacional, Mapa Topográfico Nacional de España 1:25.000, Hoja 1011-I. Guadix, Ministerio de Fomento, Madrid, 2000.

[4] F. Lima and M. Ferreira, *Minning spatial data from GPS traces for automatic road network extraction*, 6th international symposium on mobile mapping technology, Presidente Prudente, Sao Paulo, Brazil, July 21-24, 2009.

[5] X. Liu, J. Biagioni, J. Eriksson, Y. Wang, G. Forman and Y. Zhu, *Mining Large Sacale, Sparse GPS traces for Map Inference: Comparison of Approaches*, KDD12, Beijing, China, 2012.

[6] N. Schuessler and K. Axhausen, *Map-Maching of GPS traces on high resolution navigation networks using the multiple hypotesis technique (MHT)*, Working Paper 568, Institute for Transport Planning and System (IVT), ETH Zurich, Zurich, 2009.

[7] C. de Boor, *A practical guide to splines*, Springer-Verlag, New York, 2001.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Time-Aware Multi-threaded Genetic Algorithm for Accelerating a Forest Fire Spread Forecast System

Tomàs Artés[1], Andrés Cencerrado[1], Ana Cortés[1] and Tomàs Margalef[1]

[1] *Computer Architecture and Operating Systems Department,*
Universitat Autònoma de Barcelona

tomas.artes@caos.uab.es, andres.cencerrado@uab.cat, ana.cortes@uab.cat,
tomas.margalef@uab.cat

Abstract

When simulating natural hazards, data input uncertainty should be considered due to its impact into the prediction results. A way to overcome this problem consists of calibrating inaccurate input data applying computational intensive methods. However, when dealing with natural disasters, it is compulsory to provide an accurate hazard evolution forecast on time. In this paper, a multi-threaded Genetic Algorithm is proposed to exploit multi-core platform in order to accelerate a forest fire spread prediction system. The algorithm is parallelized using an hybrid MPI-OpenMP approach. The proposed solution allows to keep bounded the prediction time to the predefined time prediction requirements by including time-aware population classification, in order to allocate the most appropriate number of cores to each individual to achieve the preset deadline.

Keywords: Multi-core platforms, forest fire spread prediction, Hybrid MPI-OpenMP scheme, time assessment, core allocation

1 Introduction

A natural hazard is a possibility of a natural event that causes harm to humans. When this natural hazard causes unacceptable large numbers of fatalities and/or overwhelming property damage is a natural disaster. Wildfires are natural hazards with a high potential to become a natural disaster. For that reason, there exist a large scientific community studying such a phenomena. When dealing with an ongoing natural disaster such as a forest fire, a critical point to considering is the response time of the emergency systems and their

ability to act in the most efficient way. Experience on fire fighting and forest fire behaviour knowledge are the basic key point used to decide how to tackle an ongoing fire. In order to help fire fighting decisions, forest fire spread simulators can became a relevant tool to assess decision support systems [1]. However, to be effective, the forecasted forest fire behaviour must be delivered in advance to the predicted fire evolution. Consequently, any forest fire spread prediction system is guided by real time constraints to be useful. Furthermore, and not dismissible, there exists an inherent error related to any natural hazard prediction due to, among others, the uncertainty in the data needed to perform the forecast. For the particular case of forest fire, we can find in the literature different approaches to tackle these problems ranging from applying ensemble strategies to soften the uncertainty input parameters effects [2] to apply Kalman filter to certain input variables to tune their values [3]. Most of these approaches do not care about response time.

In this work, we focus on strategies to relieve the uncertainty effects due to the imprecision of input simulator data by ensuring a time limit. For that purpose, we used the so call Two-Stage prediction scheme, which is composed of a Calibration stage where the input parameters values are tuned to better reproduce the observed past behaviour of the fire, and those calibrated parameters are then used in the Prediction stage to forecast the forest fire evolution [4]. As a calibration strategy the Two-Stage prediction scheme uses a Genetic Algorithm (GA). However, although GAs are powerful and robust optimization techniques because of their independence of the initial guess and their few constrains on the solution domain, their main drawback is their overall run time which can easily become unacceptable. Furthermore, forest fire simulation time for a certain combination of the input parameters set, can vary from minutes to hours for the same topographic area. Consequently, it arises the need of finding a trade-off between prediction accuracy achieved thanks to the calibration strategy and the time incurred in reaching this prediction improvement. To harmonize quality and time, we propose a multi-threaded Genetic Algorithm to exploit multi-core platform in order to accelerate the Two-Stage forest fire spread prediction system. The prediction scheme has been parallelized using an hybrid MPI-OpenMP approach, where a time-aware core allocation scheme has been implemented to ensure a simulation time limit for each executed GA individual. Thus, the proposed solution allows to keep bounded the prediction time to the predefined time prediction requirements, enabling the capacity of deliver forest fire behaviour information useful to the wildfire analysts in charge of the fire management.

In the next section the hybrid MPI-OpenMP prediction framework is described. Section 3, introduces the time aware core allocation scheme. The application of the improved prediction scheme to a real case is analyzed in section 4 and, finally, the main conclusion of this work are sited in section 5.

 ISBN: 978-84-616-9216-3

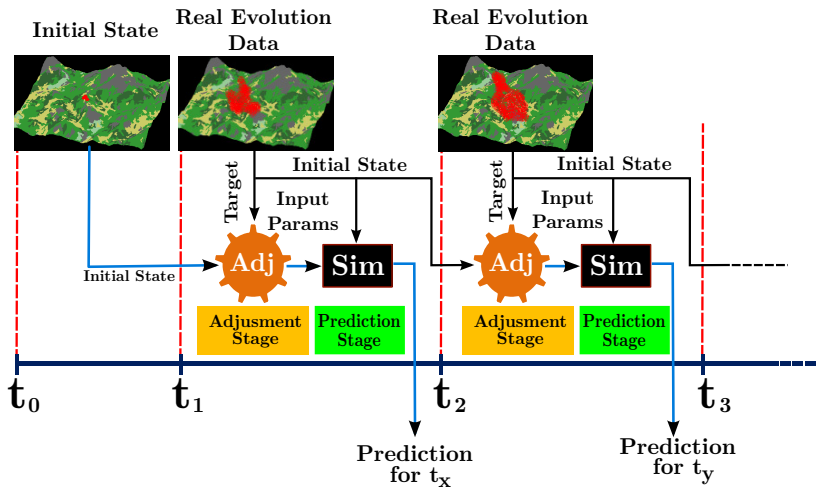


Figure 1: 2 stage prediction method

2 Hybrid MPI-OpenMP Master/Worker Prediction Scheme

A simulator independent data-driven prediction scheme is used to calibrate the input data set provided to a certain simulator [4]. For this purpose, a previous calibration step is introduced, as can be seen in figure 1. So, the input data set used for the Prediction stage is calibrated in this first stage for each Prediction step. Based on the hypothesis that the meteorological conditions will not suddenly change from the calibration stage to the prediction stage, the calibrated data set could be used to produce a more accurate prediction. Because of their outstanding results within this framework [5], this work is based on the use of Genetic Algorithm (GA) as calibration technique. The algorithm starts from an initial random population of individuals, each one representing a scenario to be simulated. An individual is composed of a set of different genes that represent input variables such as dead fuel moisture, live fuels moisture, wind speed and direction, among others.

Each individual is simulated and it is evaluated comparing the predicted and real fire propagation by estimating the fitness function (also called error function) described in equation 1. This fitness function computes *the symmetric difference between predicted and real burned areas divided by actual real spread* considering a cell based area description.

$$Difference = \frac{UnionCells - IntersectionCells}{RealCells - InitCells} \tag{1}$$

In equation 1, *UnionCells* is the number of cells which describe the surface burned considering predicted fire and the real fire. *IntersectionCells* is the number of cells burned in the real map and also in the predicted map, and *RealCells* are the cells burned in the real map. *InitCells* is the number of cells burned at the starting time. This difference takes into

account the wrong predicted burned cells (false alarms) and the real burned cells that were not predicted (misses).

According to this fitness function the whole population is ranked and the genetic operators *selection*, *elitism*, *mutation* and *crossover* are performed over the population, producing an evolved population which will have, at least, the best individual of the last generation (elitism). The new population is then evaluated in the same way. This iterative process allows us to find a good input parameter set, but it involves high computational cost due to the large amount of simulations required. Therefore, it is essential to speed up the execution keeping the accuracy of the prediction. For this reason, an implementation of the Two-Stage methodology has been developed using High Performance Computing techniques.

Since the GA fits the Master/Worker paradigm, an MPI implementation has been developed. At the first stage, the master node generates an initial random population which is distributed among the workers. Then, the workers simulate each individual and evaluate the fitness function. The errors generated by the workers are sent to the master which sorts the corresponding individuals by their error before applying the genetic operators and producing a new population.

This iterative process is repeated a fixed number of times. The last iteration (generation) contains a population from which the best individual is taken as the best solution, and then it is used in the Prediction stage.

Since every simulation can be carried out in a parallel way, the individual whose simulation takes longer determines the elapsed time for that particular generation. In order to shorten simulation times, FARSITE has been analyzed with profiling tools such as OmpP[6] and gprof [7] to determine which regions of the code could be parallelized with OpenMP. The result of such analysis determined the particular loops that could be parallelized using OpenMP pragmas. The results of such parallelization have been presented in [8]. The parallelized loops represents about 60% of one iteration execution time. It means that 40% of the iteration execution time is sequential and it implies that the speed up is not linear, but is limited by such sequential part. Figure 2 sketches the implemented Hybrid MPI-OpenMP scheme.

3 *TAC*-Two-Stage prediction scheme

One critical point when dealing with a real hazard is the response time. In order to be operative, any forecast system must release its predictions previously to the forecast event. Therefore, it is crucial to be able to keep the execution time of any prediction scheme inside a determined time bounds. In the case of the Two-stage prediction approach, the most time consuming element is the Calibration stage. This stage is an iterative process where at each iteration a wider set of forest fire spread simulations must be executed. Thus, assuming that all simulations are run using a single core (serial execution) and also assuming

 ISBN: 978-84-616-9216-3

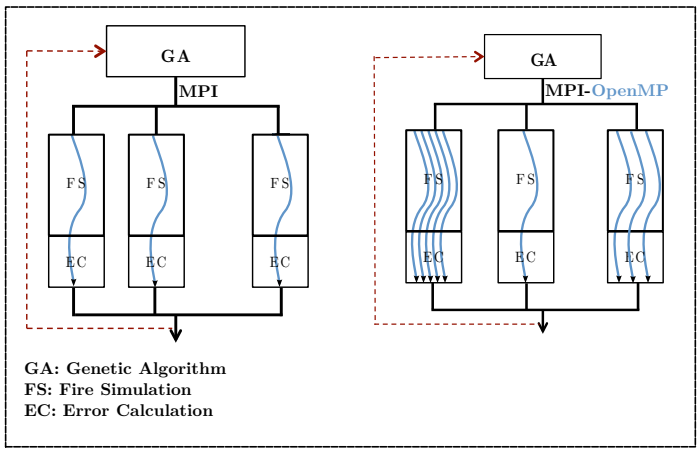


Figure 2: Hybrid MPI-OpenMP master/worker prediction scheme

that there are enough cores in the system to execute all populations members in parallel, the simulation that lasts longer will determine the duration of the corresponding iteration. When extrapolating this fact to all GA iterations, one comes out with a total calibration execution time equal to the sum of the larger simulation time existing at each iteration. Therefore the total execution time of the GA (t_{total}) could be evaluated using equation 2, where N_{Gen} is the number of generations performed of the GA, t_{Ind} stands out from the execution time of a given individual and P_{Gen} is the set of individuals (population) at the *genth* generation.

$$t_{total} = \sum_{Gen=1}^{N_{Gen}} max_{Gen}(t_{Ind}) \mid \forall Ind \in P_{Gen} \tag{2}$$

This problem is not dismissible because as it has been stated in previous works [9] the execution time of a single simulation on the same map and with the same simulation time horizon can vary from seconds to several minutes or even hours depending on the input settings of the fire spread simulator. However, if one wants to be able to deliver forest fire spread predictions within certain time ranges, we need to include to the system the ability of anticipating the simulation execution time when a certain input settings are fitted to the simulator without running that simulation. This ability would enable the capacity of limiting the total execution time of the GA by limiting the execution time of each individual iteration.

Since this knowledge is not directly available from the direct analysis of the underlying input data values, we rely on the characterization methodology described on [9]. For a given topographic area, this methodology has the ability to assess in advance the execution time of the forest fire spread simulation associated to a certain input parameter settings.

Class	Cores	Time limits
A	1	$0 < t_s \leq tmax_{Gen}$
B	2	$tmax_{gen} < t_s \leq 1.42 * tmax_{Gen}$
C	4	$1.42 * tmax_{Gen} < t_s \leq 1.81 * tmax_{Gen}$
D	8	$1.81 * tmax_{Gen} < t_s \leq 2.1 * tmax_{Gen}$

Table 1: Time limit classes for a time constraint of $tmax_{gen}$

This methodology takes advantage of the artificial intelligence field to generate a decision tree, which is able to classify the individuals of a new generated population into time classes. This classification is based on the serial version of the FARSITE simulator and, therefore, it was not able to exploit multi-core architecture. As it has been stated in the previous section, FARSITE has been parallelized using OpenMP pragmas. The parallelized part of FARSITE represents approximately 60% of the total execution time, meanwhile 40% of the execution time corresponds to a sequential part, which cannot be reduced by the implemented parallelization. Therefore, equation 3 expresses the theoretical minimum execution time we could obtain using this parallelization as a function of the number of cores (N_{Cores}), where t_s stands for the execution time of the serial version.

$$t_{par}(N_{Cores}) = 0.4 * t_s + \frac{0.6}{N_{Cores}} * ts \tag{3}$$

Furthermore, dealing with strict real time constraints implies to set up a time limit to the Calibration Stage and, consequently, to each GA iteration. Therefore, applying the FARSITE multi-threaded version and stating that the ($t_{par}(N_{Cores})$) determines the maximum GA iteration time exploiting the parallel FARSITE version, one can state the maximum serial time permitted to accomplish a generation limit time ($t_{par}(N_{Cores})$) depending on the number of cores (see equation 4).

$$t_s = \frac{t_{par}(N_{Cores}) * N_{Cores}}{0,6 + 0,4 * N_{Cores}} \tag{4}$$

Therefore, assuming a maximum GA generation time of $tmax_{Gen}$ and four available core allocation configurations: 1, 2, 4 and 8 cores per FARSITE simulation, one would be able to define four time execution classes with their respective time limits as is shown in table 1. At that point, the above mentioned characterization methodology based on decisions tree to assess in advance the serial execution time of a given FARSITE execution, could be redefined to be able to classify the individuals of a GA population according to the new time limit classes, which are associated to the number of core available to run the simulation. Therefore, the core allocation scheme become time constraint aware. That means, that the number of cores allocate to a given parallel FARSITE simulation will ensure to fit in with

 ISBN: 978-84-616-9216-3

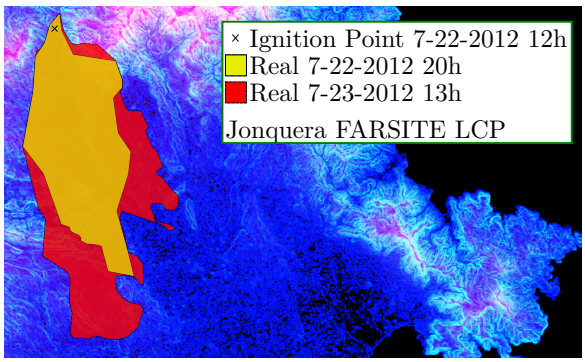


Figure 3: La Jonquera forest fire

the predetermined deadlines. This *Time Aware Classification (TAC)* approach is called *TAC-Two-Stage Prediction Scheme*. Under this scheme, any serial FARSITE simulations whose estimated prediction time goes beyonds the time limit $2.1 * tmax_{gen}$, will not be executed because there no exist any core allocation configuration that could guarantee to deliver the corresponding simulation result on time.

In the next section, the proposed *Time Aware Classification* is tested on a real case to verify the ability of this proposal to cope with the real needs of wildfire analyst during a real hazard.

4 Study case

The Mediterranean area is one of the European regions most affected for forest fires during high risk seasons. The selected study case corresponds to a region within the Mediterranean coast that is affected for forest fires almost every year. In particular, we used a fire that occurred in *La Jonquera* (North-East of Catalonia, Spain) in July 2012. This hazard devastated near 13,000ha and two people died. Figure 3 shows the burn area associated to this forest fire for two different time instants during the fire occurrence. The computing platform used to test the proposed scheme consists of two PowerEdge C6145 nodes. Each node has 4 AMD Opteron^{TM}6376 of 16 cores with 128GB of DDR3 1600 MHz.

In order to evaluate the Two-Stage prediction scheme including *TAC*, we have self-imposed a calibration time of 3 hours. In fact, this deadline is based on the fire perimeters acquisition frequency provided by the Terra and Aqua NASA's satellites after being processed by EFFIS (European Forest Fire Information System)[10].

As it has been previously described, the Calibration stage implements a GA to reduce the prediction inaccuracy due to the input data uncertainty. The GA population size has been set-up to 64 individuals and the GA has been iterated ten times. Therefore, since the

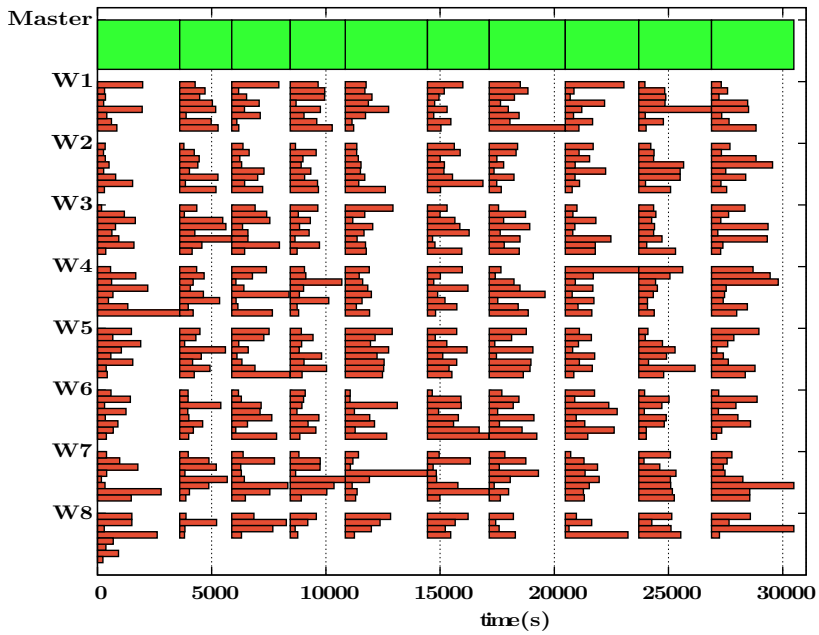


Figure 4: Execution tracing of calibration stage allocating one core per individual

calibration stage has been limited to last 10800 seconds, the maximum execution time of one GA iteration is limited to 1080 seconds. Then, assuming the best executing scenario where all GA individual are executed in parallel, the maximum serial time that should be executed at each GA iteration should last 1080 seconds. However, as it is shown in figure 4, when no time aware mechanism is introduced to the GA Calibration-stage, the unpredictable nature of GA population members, generates GA total execution time that widely exceeds the self-imposed time constraints. In particular, the total time is 30485 seconds what represents almost tree times the predefined time limit (10800 seconds). The execution trace depicted in figure 4 has been obtained from running the complete prediction system using a master/worker scheme with 8 workers, each one having eight core assigned. Therefore, the total number of cores used for this case is 64 cores and each one executes a single serial FARSITE simulation. As we can observe, this scheme is seriously penalized when one individual simulation lasts more than 1080 seconds, as it happens, for example at iteration 5. In this case, worker 7 has 8 serial simulations allocated, each one running on a single core. One of these simulations lasts 3600 seconds, limiting the corresponding GA iteration to this time. The proposed Two-Stage prediction scheme with *Time Aware Classification* (*TAC-Two Stage*) exploits the FARSITE OpenMP paralellization together with the *Time Aware Classification* described in section 3, to accomplish the predefined time constraints. In the case study, the time of each GA iteration has been set up to 1080

 ISBN: 978-84-616-9216-3

Class	Cores	Time limits
A	1	$0 < t_s \leq 1080$
B	2	$1080 < t_s \leq 1547$
C	4	$1547 < t_s \leq 1966$
D	8	$1966 < t_s \leq 2273$

Table 2: Time limit classes for a for a time constraint of 1080 seconds.

seconds. Then, assuming that the computational platform has enough cores to be able to run all GA individuals (64 individuals) in parallel, independently on the number of cores allocated to each one, the maximum parallel time of one simulation is limited by 1080 seconds. Then, substituting the values in equations 3 the resulting equation corresponds to equation 5 and, consequently, the resulting serial time depending on the number of cores allocates is the one stated in equation 6.

$$1080 = 0.4 * t_s + \frac{0.6}{N_{Cores}} * t_s \tag{5}$$

$$t_s = \frac{N_{Cores} * 1080}{0,6 + 0,4 * N_{Cores}} \tag{6}$$

Therefore, the resulting time limits from the serial times allowed to be executed depending on the number of cores used are the ones shown in table 2. Any execution of the forest fire simulator, which estimated serial execution time lasts more than 2273 seconds will be killed and, consequently, eliminated from the calibration process. In order to keep the population size constant, those killed individuals are replaced by new individuals at run time for not to penalize the GA search. The previous assumption of using enough cores to be able to run all simulations in parallel in one GA iteration, should be also accomplish when running the same experiment but using the *TAC-Two-Stage* prediction scheme. For that reason, the number of workers has been duplicated to be able to use 128 cores. In figure 5, we can see the GA evolution and the execution of the individuals for all generations taking into account the number of cores allocated to each one. As wider the line is, more cores are used for the corresponding FARSITE simulation. For example, at the first iteration, the worker labeled *worker 1* has assigned one single FARSITE simulation using 8 cores, meanwhile *worker 9* has assigned 4 FARSITE simulations running on 2 cores each. As we can observed, the self-imposed limit of 10800 seconds for the Calibration stage has been accomplished. However, one concern that could arise from this approach is directly related to the error achieved at the end of the calibration stage. Since individuals that are classified as too long (more than 2273 seconds for this particular case) are discarded and substituted for new individual, one can be upset about the loss of diversity in the GA population and how it affect to the final result. However, as it is shown in figures 6, where the error

 ISBN: 978-84-616-9216-3

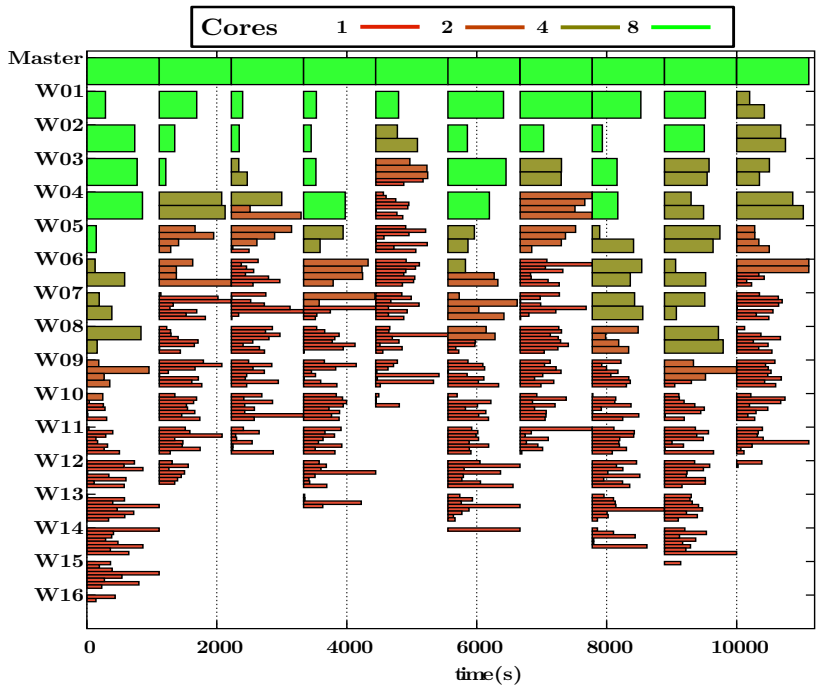


Figure 5: Execution trace of the Calibration stage using the *Time Aware Classification* (TAC-Two-Stage prediction scheme)

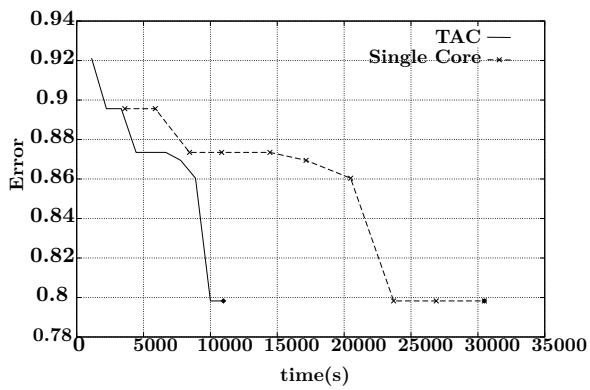


Figure 6: Error evolution through time with and without *Time Aware Classification* in the Two-Stage prediction scheme

 ISBN: 978-84-616-9216-3

evolution through time is depicted for both approaches, not using TAC and including TAC in the prediction scheme, the final result is almost the same in terms of error and, as it is desired, the time has been controlled from 30485 to 10121 seconds. Therefore, to be able to assess in advance an interval execution time (class) for any FARSITE simulation, enables the Two-Stage prediction system the capacity of classifying the GA individual according to their estimated elapsed time. This feature permits to include core allocation strategies with the aim to accomplish real time constraints. This preliminary experiments show the viability of this proposal giving light to new point to explore such as the inclusion of a resource manager, which keeps track of having a good efficiency of the system.

5 Conclusions

In this work we describe time-aware parallel forest fire spread prediction scheme. The proposed TAC-Two-Stage prediction approach exploits the implemented multi-thread FARSITE version to fit the prediction time within a present time limit. To achieve such a goal, a time-aware core allocation strategy has been included in the basic Two-Stage prediction scheme, which is able to estimate in advance the maximum execution time of a given simulation and, consequently, determine at run time, how many core allocates to each simulation to accomplish the desired time constraints. This improvement enables the system to be able to deliver prediction results under real time constraints provided by the forest fire management services. The proposed strategy has been proven using a real forest fire that took place on Catalonia on July 2012. This preliminary results denote the ability of the system to bound the total prediction time to the predefined time limit without loosing accuracy compared to the prediction results provided by the system without incorporating any time-aware strategy. The described time-aware core allocation strategy relies on the ability of having an accurate classification and, on the fact that all simulations can be executed on parallel. Current work is performed to improve efficiency by keeping time constraints when the computational platform has not enough core to ensure the whole parallelism within one GA iteration.

The time-aware core allocation policy has been implemented with the capability to change the generation time constraint during the calibration execution what enables new possibilities to the system such as apply strategies to avoid any idle time in the worker processes and incorporate the ability to dynamically either enlarge or shrink the populations size as it is needed to ensure the time constraints.

Acknowledgements

This work has been supported by MICINN-Spain under contract TIN2011-28689-C02-01 and by the Catalan government under grant 2014-SGR-576.

References

[1] M. A. Finney. *FARSITE, Fire Area Simulator–model development and evaluation.* Res. Pap. RMRS-RP-4, Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1998.

[2] Dario Rodriguez-Aseretto, Daniele de Rigo, Margherita Di Leo, Ana Cortés, and Jesús San-Miguel-Ayanz. A data-driven model for large wildfire behaviour prediction in europe. *Procedia Computer Science*, 18(0):1861 – 1870, 2013.

[3] Jan Mandel, Elhoucine Bergou, and Serge Gratton. 4dvar by ensemble kalman smoother. *arXiv preprint arXiv:1304.5271*, 2013.

[4] Baker Abdalhaq, Ana Cortés, Tomàs Margalef, and Emilio Luque. Enhancing wildland fire prediction on cluster systems applying evolutionary optimization techniques. *Future Generation Computer Systems*, 21(1):61 – 67, 2005.

[5] Mónica Denham, Kerstin Wendt, Germán Bianchini, Ana Cortés, and Tomàs Margalef. Dynamic data-driven genetic algorithm for forest fire spread prediction. *Journal of Computational Science*, 3(5):398–404, 2012.

[6] K. Fürlinger and M. Gerndt. A Profiling Tool for OpenMP. *OpenMP Shared Memory Parallel Programming*, pages 15–23, 2008.

[7] Susan L. Graham, Peter B. Kessler, and Marshall K. McKusick. gprof: a call graph execution profiler. *SIGPLAN Not.*, 39(4):49–57, April 2004.

[8] Tomàs Artés, Andrés Cencerrado, Ana Cortés, and Tomàs Margalef. Relieving the effects of uncertainty in forest fire spread prediction by hybrid mpi-openmp parallel strategies. *Procedia Computer Science*, 18:2278–2287, 2013.

[9] Andrés Cencerrado, Ana Cortés, and Tomàs Margalef. Response time assessment in forest fire spread simulation: An integrated methodology for efficient exploitation of available prediction time. *Environmental Modelling & Software*, 54:153–164, 2014.

[10] Jesus San-Miguel-Ayanz, P Barbosa, G Schmuck, G Libertà, and J Meyer-Roux. The european forest fire information system. In *AGILE 2003: 6th AGILE Conference on Geographic Information Science*, page 27. PPUR presses polytechniques, 2003. http://forest.jrc.ec.europa.eu/effis/about-effis/.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Dissolution and bulk erosion in viscoelastic materials: numerical study

E. Azhdari[1], J.A. Ferreira[1], Paula de Oliveira[1] and P.M. da Silva[2]

[1] *CMUC, Department of Mathematics, University of Coimbra - Portugal*

[2] *CMUC, Instituto Politécnico de Coimbra, ISEC, DFM, Rua Pedro Nunes, 3030-199 Coimbra, Portugal*

emails: ebrahim@mat.uc.pt, ferreira@mat.uc.pt, poliveir@mat.uc.pt, pascals@isec.pt

Abstract

A mathematical model to simulate drug delivery from a vicoelastic erodible matrix is presented in this paper. The drug is initially distributed in the matrix which is in contact with water. The entrance of water in the material changes the molecular weight and bulk erosion can be developed depending on how fast is this entrance and how fast degradation occurs. The viscoelastic properties of the matrix also change in the presence of water as the molecular weight changes. The model is represented by a system of quasi linear partial differential equations that take into account different phenomena: the uptake of water, the decreasing of the molecular weight, the viscolestic behaviour, the dissolution of the solid drug and the delivery of the dissolved drug. Numerical simulations illustrating the behaviour of the model are included.

Key words: dissolution, diffusion, molecular weight, bulk erosion, vicoelastic polymers, IMEX method

1 Mathematical model

We consider a biodegradable viscoelastic polymeric matrix, $\Omega \subseteq \mathbb{R}^2$, with boundary $\partial\Omega$ and containing a limited amount of drug. The matrix enters in contact with water and as the water diffuses into the matrix, a hydratation process, that modifies the viscoelastic properties of the polymer, takes place. The molecular weight decreases and the drug starts to dissolve.

In [13] a system that describes the sorption of water, by a loaded erodible matrix and the release of drug was proposed. However the viscoelastic properties of the matrix were not considered. In this paper we present a general model, which generalizes the model in [13], by considering the viscoelastic behaviour of the polymer (see for instance [1],[2], [6],[10], [14], [16]).

We consider a system of partial differential equations (PDE's) that describe the whole process: the entrance of water into the polymer and its consumption in the hydrolysis process; the decreasing of the molecular weight; the evolution of the stress and strain; the dissolution and the diffusion of the dissolved drug. The system reads

$$\begin{cases} \dfrac{\partial C_W}{\partial t} = \nabla \cdot (D_W \nabla C_W) + \nabla \cdot (D_v \nabla \sigma) - k C_W M & \text{in } \Omega \times (0,T], \\ \dfrac{\partial M}{\partial t} = -k C_W M & \text{in } \Omega \times (0,T], \\ \dfrac{\partial \sigma}{\partial t} + \dfrac{E(M)}{\mu(M)}\sigma = -E(M)\dfrac{\partial C_W}{\partial t} & \text{in } \Omega \times (0,T], \\ \dfrac{\partial C_S}{\partial t} = -k_{dis} C_{Sn} C_{An} C_{Wn} & \text{in } \Omega \times (0,T], \\ \dfrac{\partial C_A}{\partial t} = \nabla \cdot (D(M) \nabla C_A) + k_{dis} C_{Sn} C_{An} C_{Wn} & \text{in } \Omega \times (0,T]. \end{cases} \tag{1}$$

In (1) C_W, C_S and C_A represent the concentration of water, solid drug and dissolved drug in the polymeric matrix, respectively, M is the molecular weight of the polymer and σ is the stress response to the strain exerted by the water molecules.

The first diffusion-reaction equation of (1) describes the diffusion of water into the matrix and its consumption in the hydrolysis. In this equation D_W represents the diffusion tensor of water in the polymeric matrix. We consider an isotropic medium where the diffusion tensors are diagonal with equal diagonal elements. For example, $D_W = D_W I$, where I is the 2×2 identity matrix. The viscoelastic opposition to the water entrance is represented by $\nabla \cdot (D_v \nabla \sigma)$ where D_v is a viscoelastic diffusion tensor. This term states that the polymer acts as a barrier to the diffusion of water into the polymeric matrix. The term $-kC_W M$ represents the consumption of water in the hydrolysis of the polymer([7]).

Since the water diffuses into the polymeric matrix the molecules of water react with the polymer and the bounds between the polymeric chains are broken leading to a decrease in the molecular weight of the matrix. This process is described by the second equation of (1) ([13]).

We assume that the viscoelastic behaviour of the polymer can be modelled by Maxwell fluid model

$$\frac{\partial \sigma}{\partial t} + \frac{E}{\mu}\sigma = E\frac{\partial \epsilon}{\partial t}, \tag{2}$$

where E represents the Young modulus of the material, μ is its viscosity and ϵ is the strain produced by the water molecules. We assume that the strain and the concentration of water

 ISBN: 978-84-616-9216-3

are proportional, that is, there $k_1 > 0$ such that $\epsilon = k_1 C_W$. As the polymer acts as a barrier to the entrance of the water, then σ and ϵ are of opposite sign, and a minus sign should be considered in the right hand side of (2) ([7]).

Based on the results presented for instance in [1], [2], [6], [10], [14] and [16], we assume that the Young modulus and the viscosity depend on the molecular weight. In fact the Young modulus varies significantly in a biodegradable polymeric matrix due to the heterogeneous nature of the hydrolysis reaction that leads to the cleavages of the polymeric chains. As the degradation processes evolves, the Young modulus decreases ([12]). Moreover a functional relation between the viscosity and the molecular weight represented by Mark-Houwink equation ([11]) is applied. The expressions used to represent the behaviours of $E(M)$ and $\mu((M))$ are $E(M) = E_0 M^\alpha$ and $\mu(M) = \mu_0 M^\beta$ where E_0, μ_0, α and β are constant ([11, 12]).

The evolution in time of the solid drug is described by the fourth equation of (1) where k_{dis} is the dissolution rate, C_{Sn} is the normalized concentration of solid drug in the polymeric matrix, C_{An} is the difference between the dissolved drug concentration and its maximum solubility (C_{Amx}), normalized by C_{Amx}, C_{Wn} is the normalized concentration of water ($\frac{C_W}{C_{Wout}}$). In this last expression C_{Wout} is the concentration of water outside of the polymeric matrix. The evolution of the concentration of dissolved drug in the matrix is defined by the last equation of (1) where Fick's law and the dissolution source were taken into account.

As the degradation occurs the molecular weight decreases and the permeability of the polymer increases. This leads to an increasing of the diffusion coefficient ([15]) that can be represented by

$$D(M) = D_A e^{\bar{k}\frac{M_0 - M}{M_0}},$$

where D_A is the diffusion coefficient of the drug in the non hydrolyzed polymer, M_0 is its initial molecular weight and $\bar{k}$ is a positive constant.

System (1) is completed with the initial conditions

$$\begin{cases} C_W(0) = 0 & \text{in } \Omega, \\ \sigma(0) = \sigma_0 & \text{in } \Omega, \\ M(0) = M_0 & \text{in } \Omega, \\ C_S(0) = C_{S0} & \text{in } \Omega, \\ C_A(0) = 0 & \text{in } \Omega, \end{cases} \tag{3}$$

where σ_0 represents the initial stress of the polymer and C_{S0} is the initial concentration of solid drug in the polymeric matrix.

Degradation of the polymeric matrix can be one of the two types: surface and bulk. Surface degradation occurs because degradation is faster than the entrance of water in

 ISBN: 978-84-616-9216-3

the system. In this case the cleavage of polymeric chains occurs mainly in the outermost polymeric layers. Bulk degradation occurs when the degradation is slower than the water uptake. The entire system is rapidly hydrated and polymeric chains are cleaved through all the polymeric structure ([15]).

In what follows we assume that bulk degradation occurs and that the physical domain maintained during all diffusion process. The entrance of water occurs due to the difference of concentrations in the polymer and in the water. Then the system (1) and the initial conditions are coupled with the following boundary condition

$$\begin{cases} J \cdot \eta = A_c(C_W - C_{Wout}) & \text{on } \partial\Omega \times (0,T], \\ C_A = 0 & \text{on } \partial\Omega \times (0,T], \end{cases} \tag{4}$$

where J represents the flux defined by $J = -D_W \nabla C_W - D_v \nabla \sigma$, η is the unit outward normal to $\partial\Omega$, A_c is the permeability constant and C_{Wout} denotes the water concentration out of the polymeric matrix.

The aim of this paper is to present a numerical method to solve (1), (3) and (4) and to study the qualitative behaviour of the numerical solution. In Section 2 Implicit-Explicit method (IMEX) is introduced and its convergence is numerically studied. The qualitative behavior of the solution is analyses in Section 3. Finally in Section 4 we present some conclusions.

2 Numerical method

In this section we introduce a finite difference method to solve (1), (3), (4). Let Ω be the square $(0,L) \times (0,L)$, where L represents the thickness of the polymer. We fix $h > 0$ and we define in $\overline{\Omega}$ the grid

$$\overline{\Omega}_h = \Big\{ (x_i, y_j), i, j = 0, \ldots, N, x_0, y_0 = 0, x_N, y_N = L, \\ x_i - x_{i-1} = h, y_j - y_{j-1} = h, i, j = 1, \ldots, N \Big\}.$$

By Ω_h and $\partial\Omega_h$ we represent the mesh nodes of $\overline{\Omega}_h$ that are in Ω and on the boundary $\partial\Omega$, respectively. Let u_h and v_h be grid functions defined in $\overline{\Omega}_h$. To discretize the spatial derivatives we introduce the second order finite difference operator

$$D_x^*\big(a(v_h)D_{-x}u_h\big)(x_i,y_j) = \frac{1}{h}\Big(a(A_{h,x}v_h(x_{i+1},y_j))D_{-x}u_h(x_{i+1},y_j) - a(A_{h,x}v_h(x_i,y_j))D_{-x}u_h(x_i,y_j)\Big),$$

where D_{-x} denotes the backward finite difference operator with respect to the x-variable and $A_{h,x}$ is the following average operator

$$A_{h,x}v_h(x_\ell, y_j) = \frac{1}{2}\Big(v_h(x_\ell, y_j) + v_h(x_{\ell-1}, y_j)\Big).$$

 ISBN: 978-84-616-9216-3

The finite difference operator $D_y^*\big(b(v_h)D_{-y}u_h\big)(x_i,y_j)$ is defined analogously considering the backward finite difference operator with respect to the y-variable, D_{-y}, and the average operator $A_{h,y}$. If B is a diagonal matrix with entries a and b we use the following notation

$$\nabla_h^*.\big(B(v_h)\nabla_h u_h\big) = D_x^*\big(a(v_h)D_{-x}u_h\big) + D_y^*\big(b(v_h)D_{-y}u_h\big).$$

In $[0,T]$ we consider the following time grid

$$\Big\{t_n, n=0,\ldots,M_{\Delta t}, t_0=0, t_{M_{\Delta t}}=T, t_n-t_{n-1}=\Delta t, n=1,\ldots,M_{\Delta t}\Big\}.$$

By D_{-t} we denote the backward finite difference operator with respect to the variable t. Let $p_h^n(x_i,y_j)$ stands for an approximation of $p(x_i,y_j,t_n)$.

To solve numerically the initial boundary value problem (1), (3), (4) we consider the **IMEX** method defined by

$$\begin{cases} D_{-t}C_{W,h}^{n+1} = \nabla_h^*.\Big(D_W\nabla_h C_{W,h}^{n+1}\Big) + \nabla_h^*.\Big(D_v\nabla_h\sigma_h^n\Big) - kC_{W,h}^nM_h^n \ \text{ in}\,\Omega_h \\ D_{-t}M_h^{n+1} = -kC_{W,h}^{n+1}M_h^n \ \text{ in}\,\overline{\Omega}_h \\ D_{-t}\sigma_h^{n+1} + \dfrac{E_0(M_h^{n+1})^\alpha}{\mu_0(M_h^{n+1})^\beta}\sigma_h^n = -E_0(M_h^{n+1})^\alpha D_{-t}C_{W,h}^{n+1} \text{ in } \overline{\Omega}_h \\ D_{-t}C_{S,h}^{n+1} = -\dfrac{k_{dis}}{C_{S0}C_{Amx}C_{Wout}}C_{S,h}^n(C_{Amx}-C_{A,h}^n)C_{W,h}^{n+1} \ \text{ in}\,\overline{\Omega}_h \\ D_{-t}C_{A,h}^{n+1} = \nabla_h^*.\Big(D(M_h^{n+1})\nabla_h C_{A,h}^{n+1}\Big) + \dfrac{k_{dis}}{C_{S0}C_{Amx}C_{Wout}}C_{S,h}^{n+1}\Big(C_{Amx}-C_{A,h}^n\Big)C_{W,h}^{n+1} \ \text{ in}\,\Omega_h \end{cases} \tag{5}$$

for $n=0,\ldots,M_{\Delta t}-1$,

$$\begin{cases} C_{W,h}^0 = 0 \ \text{ in}\,\Omega_h \\ \sigma_h^0 = \sigma(0) \ \text{ in}\,\Omega_h \\ M_h^0 = M(0) \ \text{ in}\,\Omega_h \\ C_{S,h}^0 = C_S(0) \ \text{ in}\,\Omega_h \\ C_{A,h}^0 = 0 \ \text{ in}\,\Omega_h \end{cases} \tag{6}$$

and

$$\begin{cases} J_h^{n+1}.\eta = A_c(C_{W,h}^{n+1}-C_{Wout}) \ \text{ on}\ \partial\Omega_h \\ C_{A,h}^{n+1} = 0 \ \text{ on}\ \partial\Omega_h, \end{cases} \tag{7}$$

where

$$J_h^{n+1} = -D_WD_\eta C_{W,h}^{n+1} - D_vD_\eta\sigma_h^n,$$

 ISBN: 978-84-616-9216-3

and D_η is the boundary operator

$$D_\eta v_h(x_i, y_j) = \begin{cases} -D_x v_h(x_0, y_j), & i = 0 \\ D_{-x} v_h(x_N, y_j), & i = N \\ -D_y v_h(x_i, y_0), & j = 0 \\ D_{-y} v_h(x_i, y_N), & j = N \end{cases}$$

for $(x_i, y_j) \in \partial\Omega_h$.

3 Qualitative behaviour of the model

In this section we illustrate the influence of the parameters on the behaviour of the model. The values of the parameters are present in Table 1 and some of them were obtained from [13]. We start by analyzing numerically the convergence properties of the numerical scheme.

Parameter	Value	Parameter	Value
D_A	5.94×10^{-2}	E_0	1×10^{-3}
D_v	2×10^{-2}	μ_0	1×10^{-1}
D_W	4.61×10^{-2}	k_{dis}	4.6×10^{-2}
k	1×10^{-2}	M_0	8.3×10^{-2}
σ_0	5×10^{-2}	C_{Wout}	5.55×10^{-1}
C_{Amx}	2.184×10^{-2}	A_c	1×10^{-2}
C_{S0}	288.42×10^{-2}	α	0.2
β	0.7	L	1
Δt	1×10^{-4}	h	0.01

Table 1: Parameter values used for the simulation.

Table 2 contains the errors for C_W and C_A defined by

$$Error(C) = \max_{n=1,\ldots,M_{\Delta t}} \max_{\Omega_h} |C_h^n - \overline{C}_h^n|,$$

where $C = C_W, C_A$ and $\overline{C}_h^n$ is a reference solution obtained with a fine grid defined by $\Delta t = 10^{-5}$ and $h = 0.001$.

h	$Error(C_W)$	$Error(C_A)$
0.01	0.0048	5.1432×10^{-8}
0.005	0.0032	4.8043×10^{-8}
0.004	0.0029	4.4917×10^{-8}
0.002	0.0017	2.9373×10^{-8}

Table 2: Errors for different step-sizes in space.

 ISBN: 978-84-616-9216-3

The results of Table 2 suggest the convergence of the IMEX method.

Let the mass of water and drug, inside the matrix, be defined by

$$\mathcal{M}_i(t) = \int_\Omega C_i(t) dxdy,$$

where $i = W, A$, for $t \in [0, T]$. A numerical approximation for $M_i(t)$ is computed with the trapezoidal rule.

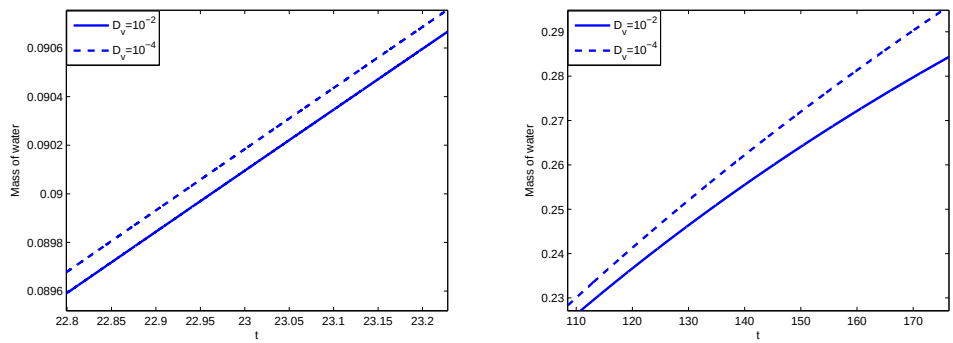


Figure 1: Influence of D_v on the mass of the water.

In Figure 1 we plot the dependence on the viscoelastic diffusion coefficient D_v of the mass of water. We observe that the polymer acts as a barrier to the entrance of water into the polymer. In other words, the non Fickian flux $-D_v \nabla \sigma$ decreases the Fickian flux, $-D_W \nabla C_W$. According to this description an increase in D_v leads to a decrease of M_W.

The influence of the Young modulus E on M_W is presented in Figure 2 (left). near $t = 2$. It is well known that the crosslink density of the polymer is proportional to the Young modulus E. Consequently as this constant increases the resistance of the polymer to the entrance of water also increases leading to a decreasing of the mass of water.

The influence of the polymer degradation rate, k, is presented in Figure 2 (right). As expected, if the degradation rate increases, then the delivery rate of the dissolved drug also increases.

The behaviour of the mass of dissolved drug is presented in Figure 3, for different thickness of the polymer. We observe that the maximum value of the mass of dissolved drug in thinner polymers is higher and less time is required to achieve this maximum.

In Figure 4 the mass of water inside the polymer, for different values of L, is plotted. In the thinner polymer more time is required for the mass to reach the steady state. We also observe that the value of the steady state in the polymer with $L = 0.1$ is 0.0555 while in the polymer with $L = 0.5$ is 0.2769.

 ISBN: 978-84-616-9216-3

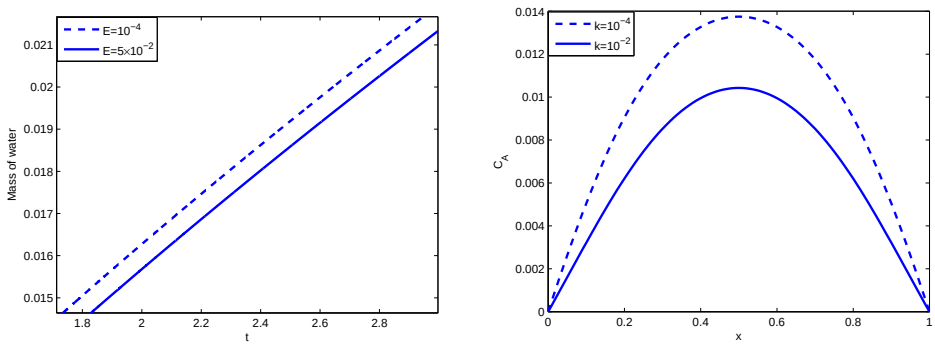


Figure 2: Mass of water for different E's (left); concentration of dissolved drug C_A for different k (right).

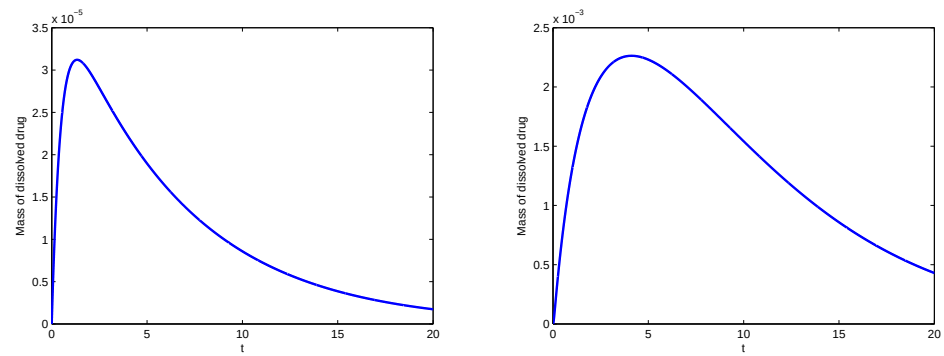


Figure 3: Mass of dissolved drug inside the polymer with $L = 0.1$ (left) and $L = 0.5$ (right).

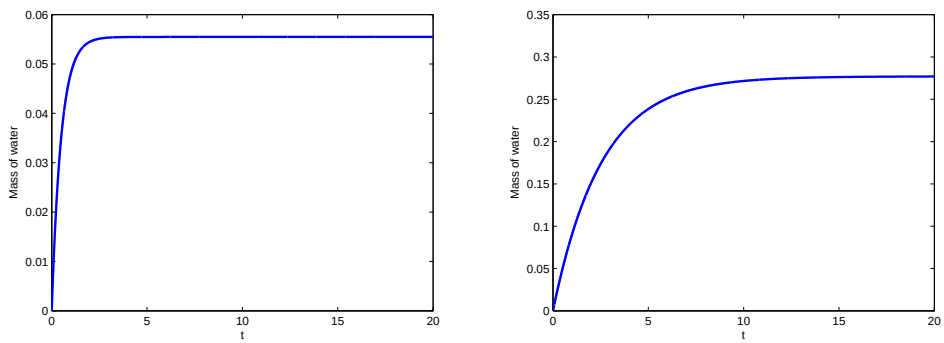


Figure 4: Mass of water inside the polymer with $L = 0.1$ (left) and $L = 0.5$ (right).

 ISBN: 978-84-616-9216-3

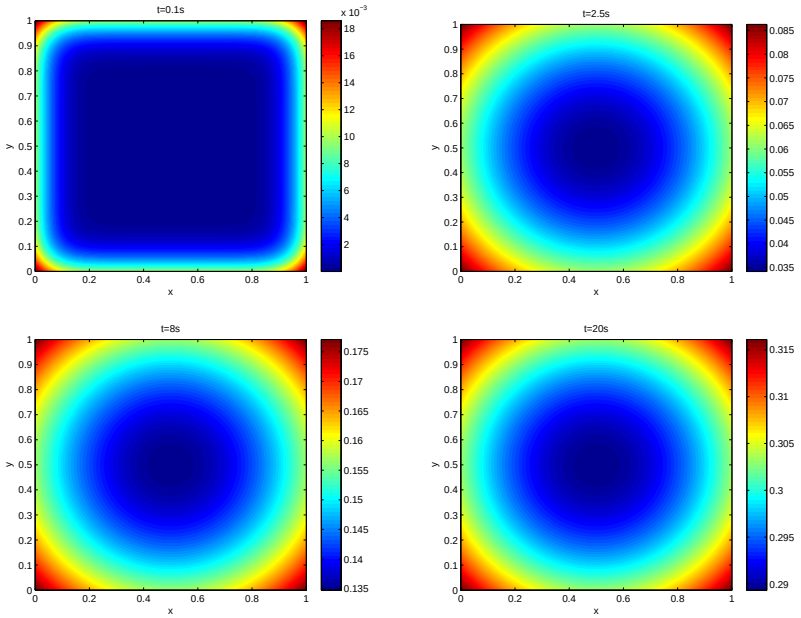


Figure 5: Concentration of water for different times.

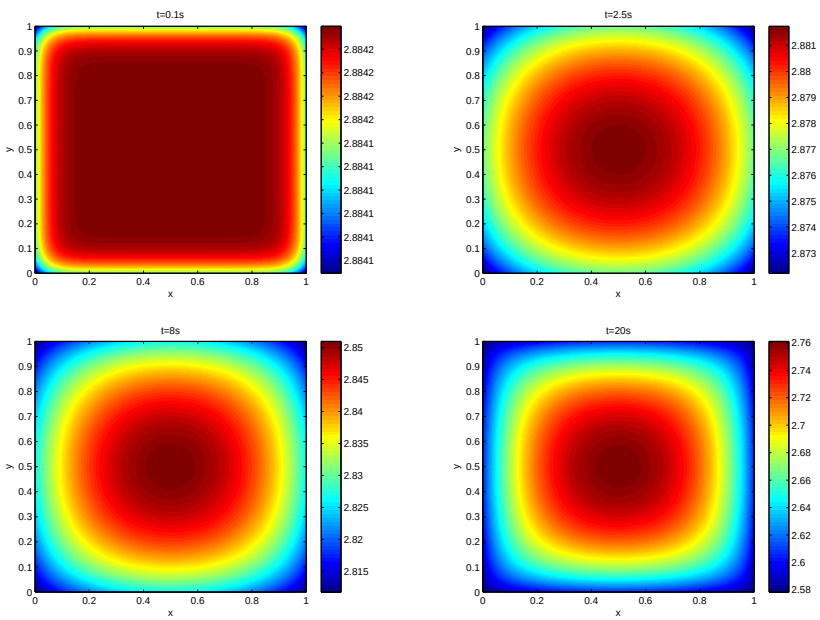


Figure 6: Concentration of solid drug for different times.

 ISBN: 978-84-616-9216-3

Figure 5 illustrates the behavior of the concentration of water into the polymeric matrix at different times. We observe that the concentration increases as time increases and the behavior is homogeneous since the diffusion coefficient is constant.

The concentration of solid drug and dissolved drug, respectively, at different times are shown in Figures 6 and 7. Te regions where the concentration of water is higher, correspond to regions where the concentration of solid drug is lower. We also note that when the concentration of solid drug decreases, the concentration of dissolved drug increases.

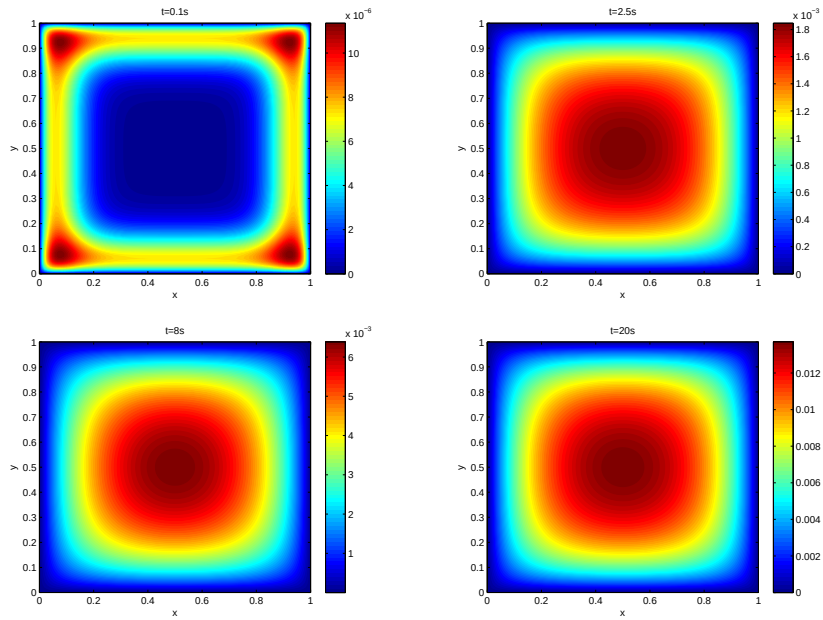


Figure 7: Concentration of dissolved drug for different times.

4 Conclusions

In this paper we describe a process of sorption of a solvent by a biodegradable polymeric matrix, when bulk erosion occurs, and the simultaneous release of a drug. Numerical results that highlight the whole process are presented. These results are physically sound. The influence of the crosslinking density of the polymer is shown to delay the drug release. In fact a larger Young modulus exerts a larger opposition to the solvent penetration. Bulk erosion which is governed by the degradation rate speeds up the release of drug. The dependence on the dimensions of the matrix is also illustrated.

The theoretical study of the initial boundary value problem (1), (3) and (4) will be object of a future work. We intent also to analyse the occurrence of surface degradation.

 ISBN: 978-84-616-9216-3

Acknowledgements

This work was partially supported by the Centro de Matemática da Universidade de Coimbra (CMUC), funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e Tecnologia under the project PEst-C/MAT/UI0324/2013.

References

[1] R. Al-Itry, K. Lamnawar, A. Maazouz , *Improvement of thermal stability, rheological and mechanical properties of PLA, PBAT and their blends by reactive extrusion with functionalized epoxy*, Polym Degrad Stab **97** (2012) 1898–1914.

[2] M. Al-Nasassrah, F. Podczeck, J.Newton, *The effect of an increase in chain length on the mechanical properties of polyethylene glycols*, Eur. J. Pharm. Biopharm. **46** (1998) 31–38.

[3] E. Azhdari, J.A. Ferreira, P. de Oliveira, P.M. da Silva, *Analytical and numerical study of diffusion through biodegradable viscoelastic materials*, Proceedings of the 13th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2013, **I**, (2013), 174–184.

[4] E. Azhdari, J.A. Ferreira, P. de Oliveira, P.M. da Silva, *Drug delivery from an ocular implant into the vitreous chamber of the eye*, Proceedings of the 13th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2013, **I**, (2013), 185–195.

[5] E. Azhdari, J. A. Ferreira, P. de Oliveira, P. M. da Silva, *Diffusion, viscoelasticity and erosion: analytical study and medical applications*, J. Comput. Appl. Math. (2014) doi: 10.1016/j.cam.2014.01.025.

[6] S. Baruah, N. Laskar, *Relation between molecular weight and viscosity for Polydispersed Pol(n-docosyl acrylate)*, Polym. J. **28** (1996) 893–895.

[7] J.A. Ferreira, M. Grassi, E. Gudiño, P. de Oliveira, *A 3D model for mechanistic control drug release*, SIAM J. Appl. Math., 74 (2014) 620633. .

[8] A.Gopferich, R. Langer, *Modeling of polymer erosion*, Macromolecules **26** (1993) 4105–4112.

[9] A. Gopferich, *Mechanisms of polymer degradation and erosion*, Biomaterials **17** (1996) 103–114.

[10] A. Izuka, H. Winter, T. Hashimoto, *Molecular weight dependence of viscoelasticity of polycaprolactone critical gels*, Macromolecules **25** (1992) 2422–2428.

[11] S. Luo, D. T. Grubba and A. N. Netravali, *The effect of molecular weight on the lamellar structure, thermal and mechanical properties of poly (hydroxybutyrate-co-hydroxyvalerates)*, Polymer **43** (2002) 4159–4166.

[12] Y. Wang, X. Han, J. Pan, C. Sinka, *An entropy spring model for the Young's modulus change of biodegradable polymers during biodegradation*, J. Mech. Behav. Biomed. **3** (2010) 14–21.

[13] S. N. Rothstein, W. J. Federspiel, S. R. Little, *A unified mathematical model for the prediction of controlled release from surface and bulk eroding polymer matrices*, Biomaterials, **30** (2009) 1657–1664.

[14] T. Rushing, R. Hester, *Intrinsic viscosity dependence on polymer molecular weight and fluid temperature*, J. Appl. Polym. Sci. **89** (2003) 2831-2835.

[15] J. Siepmann, A. Göpferich, *Mathematical modeling of bioerodible polymeric drug delivery systems*, Adv. Drug. Deliver Rev. **48** (2001) 229–247.

[16] J. Torres, C. Stafford, B. Vogt, *Impact of molecular mass on the elastic modulus of polystyrene thin films*, Polymers **51** (2010)4211–4217.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

A comparison of interval estimation methods in partially non-regular log-exponential models

I. Barranco-Chamorro[1], M.D. Jiménez-Gamero[1] and M.V. Alba-Fernández[2]

[1] *Department of Statistics and Operations Research, University of Seville (Spain)*

[2] *Department of Statistics and Operations Research, University of Jaen (Spain)*

emails: chamorro@us.es, dolores@us.es, mvalba@ujaen.es

Abstract

In this paper we study and compare the performance of approximate confidence intervals for a given parametric function based on different asymptotic approaches when sampling from partially non-regular log-exponential models. Specifically, we consider Wald-Type, Score and Likelihood-Ratio-Test intervals.

Key words: asymptotic, interval estimation, log-exponential models, partially non-regular
MSC 2000: 62F12, 62N02

1 Introduction

In this paper we consider *partially non-regular log-exponential models*. On the one hand, the term *partially non-regular* was proposed by Dubinin and Vardeman [3] to deal with parametric models involving regular and non-regular estimators for inferential purposes. On the other hand, the term *log-exponential* models refers to positive random variables such that the logarithm transformation allows us to apply results from the two-parameter exponential distribution. In this setting the point is that, unlike regular models extensively studied in the literature, the mixture of regular and non-regular estimators leads to complications with respect to inferential purposes. Recently, Barranco-Chamorro and Jiménez-Gamero [2] got approximations to the moments and the different possibilities for the limiting distributions of the maximum likelihood estimator (MLE) of a given parametric function when sampling from these distributions. In this work we study other asymptotic approaches to

deal with inferential issues in these models. Specifically, we consider Wald-Type, Score and Likelihood-Ratio-Test methods to obtain asymptotic intervals of a given parametric function. The performance of these methods is compared through simulations.

Next we introduce some notation and a brief summary of results in the two-parameter exponential distribution. Our results are also applicable to the Pareto and Power-Function distributions.

Definition 1. A random variable X follows a two-parameter exponential distribution, $E(\theta, \sigma)$, if its probability density function is

$$f(x) = \frac{1}{\sigma} e^{-(x-\theta)/\sigma} I_{[\theta,+\infty)}(x), \qquad (\theta, \sigma) \in \Theta = I\!R \times I\!R^{+}. \tag{1}$$

θ is a location parameter that also determines the support of the distribution and σ is a scale parameter.

Given a simple random sample of size n from (1), the MLEs of θ and σ are $\hat{\theta}_n = X_{(1)} = \min_{1\leq i\leq n}\{X_i\}$ and $\hat{\sigma}_n = \bar{X} - X_{(1)}$.

$\hat{\theta}_n$ is non-regular for inferential purposes whereas $\hat{\sigma}_n$ is regular.

For a real-valued function of the MLEs, h, sufficiently smooth, $h(\hat{\theta}_n, \hat{\sigma}_n)$, we got the limiting distribution (ld) of $h(\hat{\theta}_n, \hat{\sigma}_n)$ in [2]. It is given in the following theorem.

Theorem 1. (Limiting distributions of $h(\hat{\theta}_n,\ \hat{\sigma}_n)$.) Let h be a parametric function admitting a first order Taylor expansion at $(\theta, \sigma) \in \Theta$. If $D_2h(\theta, \sigma) \neq 0$ at (θ, σ) then

$$\frac{\sqrt{n}}{\hat{\tau}_n}\left\{h(\hat{\theta}_n, \hat{\sigma}_n) - h(\theta, \sigma)\right\} \xrightarrow{\mathcal{L}} Z, \quad \text{where } Z \sim N(0, 1), \tag{2}$$

$D_2h(\theta, \sigma)$ denotes the first partial derivative of h with respect to σ evaluated at a given $(\theta, \sigma) \in \Theta$, i.e. $D_2h(\theta, \sigma) = (\partial/\partial\sigma)h(\theta, \sigma)$, and $\hat{\tau}_n$ is a consistent estimator of $\sigma D_2h(\theta, \sigma)$.

2 Methods

In this section we study and compare the performance of approximate confidence intervals for a given parametric function based on different asymptotic approaches. Specifically, we consider Wald-Type, Score and Likelihood-Ratio-Test intervals.

Wald-type intervals From result given in Theorem 1, this method proposes the following $100(1-\gamma)\%$ *asymptotic confidence interval* for $h(\theta, \sigma)$

$$\left(h(\hat{\theta}_n,\ \hat{\sigma}_n) - z_{1-\gamma/2}\frac{\hat{\tau}_n}{\sqrt{n}},\ h(\hat{\theta}_n,\ \hat{\sigma}_n) + z_{1-\gamma/2}\frac{\hat{\tau}_n}{\sqrt{n}}\right), \tag{3}$$

where $\hat{\tau}_n = \hat{\sigma}_n|D_2h(\hat{\theta}_n, \hat{\sigma}_n)|$ and $z_{1-\gamma/2}$ denotes the $(1-\gamma/2)$th quantile of the N(0,1) distribution.

We highlight that (4) is easy to apply.

 ISBN: 978-84-616-9216-3

Score Intervals Let us denote by $\hat{h}_n = h(\hat{\theta}_n,\ \hat{\sigma}_n)$. Note that Wald-type intervals proposed in (3) are based on

$$\left|\frac{\hat{h}_n - h}{\widehat{se}(\hat{h}_n)}\right| \leq z_{1-\alpha/2} \tag{4}$$

where $\widehat{se}(\hat{h}_n)$ is the *estimated* standard error of $\hat{h}_n$.

The *score method* proposes to use the standard error of $\hat{h}_n$, $se(\hat{h}_n)$, instead of the estimated standard error of $\hat{h}_n$, $\widehat{se}(\hat{h}_n)$. In this way, the use of Slutsky lemma is avoided in Theorem 1 to get approximate confidence intervals. The basis of this method is given in the following theorem.

Theorem 2. If $D_2h(\theta,\sigma) \neq 0$ at (θ,σ) then

$$\frac{\sqrt{n}}{\sigma D_2 h(\theta,\sigma)}\left\{h(\hat{\theta}_n,\hat{\sigma}_n) - h(\theta,\sigma)\right\} \xrightarrow{\mathcal{L}} Z, \quad \text{where } Z \sim N(0,1). \tag{5}$$

The score interval involves the solution in h of

$$\left|\frac{\hat{h}_n - h}{se(\hat{h}_n)}\right| \leq z_{1-\alpha/2} \tag{6}$$

Intervals based on Likelihood Ratio Testing (LRT) This method considers a hypothesis test of the form

$$\begin{aligned} H_0 &: \ h(\theta,\sigma) = h_0, \qquad h_0 \in I\!R \\ H_1 &: \ h(\theta,\sigma) \neq h_0 \end{aligned}$$

The LRT statistic is given by

$$\Lambda_h = \frac{\sup\limits_{(\theta,\sigma)} \prod\limits_{i=1}^{n} f_{(\theta,\sigma)}(X_i)}{\sup\limits_{(\theta,\sigma)\text{with } h(\theta,\sigma)=h_0} \prod\limits_{i=1}^{n} f_{(\theta,\sigma)}(X_i)}$$

In order to apply this method, we need to know the *MLE* of (θ,σ) *under the constraint* proposed in H_0, $h(\theta,\sigma) = h_0$. This MLE is denoted by $(\hat{\theta}_h, \hat{\sigma}_h)$.

In this context, the question is: Which is the limiting null behaviour of $\lambda_h = 2\ \ln \Lambda_h$? The novel result is

$$\lambda_h \xrightarrow{\mathcal{L}} \chi^2_k, \qquad k = 1, 2. \tag{7}$$

That is, the limiting distribution is a chi-square distribution whose degrees of freedom, k, depends on the gradient vector of h, ∇h, at (θ_0,σ_0), with (θ_0,σ_0) a vector of parameters

satisfying the constraint proposed in H_0, $h(\theta_0, \sigma_0) = h_0$. It is possible to invert the LRT to do interval estimation. So by inverting (7), we have that

$$\left\{h_0 : \quad \lambda_h \leq \chi^2_{k,\, 1-\alpha}\right\} \tag{8}$$

is an approximate $(1-\alpha)$ confidence set for the parametric function $h(\theta, \sigma)$.

We highlight that the problem proposed in (7) and (8) is not obvious.

3 Applications and Conclusions

Applications of these models can be seen in [4] and [2]. We highlight: lifetime distributions of interest in engineering, income distributions, population statistics, and for modelling some characteristics of IP traffic in the Internet. As for parametric functions of practical interest we give results for the estimation of quantiles, Lorenz curve and Gini index when sampling from these models.

Finally, we point out that other applications can be carried out. In order to apply these results the condition $D_2 h(\theta, \sigma) \neq 0$ is extremely important so that the proposed approximations are good.

Acknowledgements

This research has been partially supported by grant UJA2013/08/01.

References

[1] M. Akahira and K. Takeuchi, *Non-regular Statistical Estimation. Lecture Notes in Statistics 107*, Springer, New York, 1995.

[2] I. Barranco-Chamorro and M. D. Jiménez-Gamero, *Asymptotic results in partially non-regular log-exponential distributions*, J. Statist. Comput. Simul. **82 (3)** (2012) 445-461.

[3] T. M. Dubinin TM and S. B. Vardeman, *Likelihood-Based Inference in Some Continuous Exponential Families with Unknown Threshold Parameters*, JASA **98 (463)** (2003) 741-749.

[4] J. F. Lawless, *Statistical Models and Methods for Lifetime Data (2nd edition)*, Wiley, Hoboken, 2003.

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

On spline-based differential quadrature

D. Barrera[1], P. González[1], F. Ibáñez[2] and M. J. Ibáñez[1]

[1] *Department of Applied Mathematics, University of Granada, 18071-Granada, Spain*

[2] *Grupo Sacyr-Vallehermoso, Paseo de la Castellana, 83, 28046-Madrid, Spain*

emails: dbarrera@ugr.es, prodelas@ugr.es, fibanez@gruposyv.com, mibanez@ugr.es

Abstract

In the paper *A general spline differential quadrature method based on quasi-interpolation*, J. Comput. Appl. Math. (2014), http://dx.doi.org/10.1016/j.cam.2014.02.006, a boolean sum differential quadrature method (DQM) was proposed by combining an spline interpolation operator having a fundamental function with minimal compact support and a spline quasi-interpolation operator exact on the space of polynomials reproduced by the first one. It is a general framework that provides results that differ from the ones obtained by defining specific schemes with a structure which depends on the degree of the B-spline to be considered. The main drawback of these boolean sum DQMs is that the number of evaluation points increases fastly with the degree of the B-spline due to the use of a quasi-interpolation operator. In this communication we propose a different construction avoiding this problem and derive explicit results for low degree B-splines.

Key words: differential quadrature, B-spline, interpolation

1 Spline differential quadrature method based on quasi-interpolation

In [1] a general spline Quadrature Differential Method (QDF) was proposed by combining interpolation and quasi-interpolation. If M_n denotes the B-spline of order $n \geq 2$ centered at the origin, the (compactly supported) fundamental function L_n of the interpolation operator $\mathcal{L}$ has the form

$$L_n = \sum_{j \in J} c_j M_n \left(2 \cdot -j\right)$$

 ISBN: 978-84-616-9216-3

for some $c_j \in \mathbb{R}$, J being a finite subset of $\mathbb{Z}$. It satisfies the interpolation conditions $L_n(j) = \delta_{j,0}$, $j \in \mathbb{Z}$, δ being the Kronecker sequence.

Since the Laurent polynomials $\Phi_k(z) := \sum_{j\in\mathbb{Z}} M_n(2j+k)\, z^{2j}$, $k = 0, 1$, have no common zeros in $\mathbb{C}\setminus\{0\}$, it follows that (see [4]) any finite sequence c satisfying the identity

$$\Phi_0 \sum_{j\in\mathbb{Z}} c_{2j} e_{2j} + \Phi_1 \sum_{j\in\mathbb{Z}} c_{2j+1} e_{2j+1} = 1$$

provides such a function L_n. The notation $e_0(z) := 1$, and $e_k(z) := z^k$, $k \geq 1$ is used. The following result on symmetric functions L_n with a small support was proved in [1].

Proposition 1 *For each $n \geq 4$, let $J := \{-d_n, \ldots, d_n\}$ where*

$$d_n := \begin{cases} \lfloor r \rfloor - 2, & \text{for } n \text{ even,} \\ \lfloor r \rfloor - 1, & \text{for } n \text{ odd,} \end{cases}$$

and $\lfloor r \rfloor$ denotes the integer part of $r \in \mathbb{R}$. Then, there are coefficients a_j, $0 \leq j \leq 2d_n$ such that the function

$$L_n = a_0 M_n(2\cdot + d_n) + \cdots + a_{2d_n} M_n(2\cdot - d_n)$$

satisfies the interpolation conditions

$$L(j) = \delta_{j,0},\ j \in \mathbb{Z}.$$

It follows that

$$\operatorname{supp} L_n \subset \begin{cases} \left[-\frac{n}{2}+1, \frac{n}{2}-1\right], & \text{for } n \text{ even,} \\ \left[-\frac{n}{2}+\frac{3}{4}, \frac{n}{2}-\frac{3}{4}\right], & \text{for } n \text{ odd.} \end{cases}$$

For a given function f defined on the real line, the spline

$$\mathcal{L}_n(f) := \sum_{i\in\mathbb{Z}} f(i)\, L(\cdot - i)$$

interpolates f at the integers. The operator $\mathcal{L}_n$, as well as its scaled version $\mathcal{L}_{n,h}(f) := \sum_{i\in\mathbb{Z}} f(hi)\, L\left(\frac{\cdot}{h} - i\right)$, does not reproduce the space $\mathbb{P}_{n-1}$ of the polynomials of degree at most $n-1$. So, in order to obtain QDMs useful in practice, it is necessary to increase the polynomial reproduction of the operator $\mathcal{L}_n$, and this is achieved by considering discrete quasi-interpolation based on M_n. A such discrete quasi-interpolant $\mathcal{Q}_n f$ for a given function f is a linear combination of integer translates of M_n. It can be written as

$$\mathcal{Q}_n f = \sum_{i\in\mathbb{Z}} f(i)\, q_n(\cdot - i),$$

 ISBN: 978-84-616-9216-3

with $q_n := \sum_{j=-m}^{m} \gamma_j M_n(\cdot - j)$ for some coefficients γ_j such that $\mathcal{Q}p = p$ when $p \in \mathbb{P}_{n-1}$. In general, the compactly supported function q_n does not satisfy the interpolation conditions.

Let us suppose that $\mathcal{Q}_n$ is a quasi-interpolation operator exact on $\mathbb{P}_{n-1}$. Then, the order of approximation of $\mathcal{L}_n$ is increased by forming the boolean sum $\mathcal{L}_n \oplus \mathcal{Q}$ of $\mathcal{L}_n$ and $\mathcal{Q}_n$, defined by the expression $\mathcal{L}_n \oplus \mathcal{Q}_n = \mathcal{L}_n + \mathcal{Q}_n - \mathcal{L}_n\mathcal{Q}_n$. This new operator inherits the interpolation properties of $\mathcal{L}_n$, the exactness of $\mathcal{Q}_n$, and produces $C^{n-2}(\mathbb{R})$ functions.

Therefore, the interpolation operator $\mathcal{I}_n := \mathcal{L}_n + \mathcal{Q}_n - \mathcal{L}_n\mathcal{Q}_n$ was used to derive new QDMs.

In order to define the interpolation operator $\mathcal{I}_n$, we need to choose a quasi-interpolation operator $\mathcal{Q}_n$ exact on $\mathbb{P}_{n-1}$. We will adopt a classical procedure (see [5]) to derive the unique quasi-interpolation operator whose coefficient linear form μ_n uses uniquely the knots iniside the support of M_n. The starting point is the expansion (see [6, p. 15])

$$s(x) = \sum_{k\in\mathbb{Z}} c_k M_n(x-k)$$

in terms of the integer translates of the B-spline M_n of a arbitrary spline s with integer of half-integer knots depending on the order of the spline. The coefficients c_k are linear forms involving the central factorial numbers (cfn for short) of the first kind. Explicitly (see [3, equality (6.2.7)]),

$$c_k = \sum_{j=0}^{n-1} \frac{(n-1-j)!}{(n-1)!} t(n, n-j)\, s^{(j)}(k),$$

where the cfn $t(n,k)$ are defined as follows (see [3, p. 421]):

$$x^{[n]} := \sum_{k=0}^{n} t(n,k)\, x^k,$$

with

$$x^{[0]} := 1,\ x^{[1]} := x, \quad \text{and} \quad x^{[n]} := x \prod_{k=1}^{n-1} \left(x + \frac{n}{2} - k\right),\ n \geq 2.$$

Equalities for s and c_k lead to the differential quasi-interpolation operator exact on $\mathbb{P}_{n-1}$

$$\mathcal{D}_n(f) := \sum_{i\in\mathbb{Z}} \left(\sum_{j=0}^{n-1} \frac{(n-1-j)!}{(n-1)!} t(n, n-j)\, f^{(j)}(i) \right) M_n(\cdot - i),$$

and the relationship between derivatives and central differences (see Proposition 6.1.1 in

 ISBN: 978-84-616-9216-3

[3]) produces the following discrete quasi-interpolation operator:

$$\mathcal{Q}_n(f) := \sum_{i\in\mathbb{Z}} \left(\sum_{j=0}^{n-1} \frac{(n-1-j)!}{(n-1)!} t(n,n-j) \left(j! \sum_{k=j}^{n-1} \frac{1}{k!} \delta^k f(i)\, t(k,j) \right) \right) M_n(\cdot - i)$$
$$= \sum_{i\in\mathbb{Z}} \left(\sum_{j=0}^{n-1} \frac{t(n,n-j)}{\binom{n-1}{j}} \left(\sum_{k=j}^{n-1} \frac{\delta^k f(i)}{k!} t(k,j) \right) \right) M_n(\cdot - i).$$

This was the quasi-interpolation operator used in combination with $\mathcal{L}_n$ to approximate the derivatives in the QDM.

2 Spline based Differential Quadrature without quasi-interpolation

The main drawback of these boolean-sum based DQMs is that the number of evaluation points in the expressions that approximate the derivatives at the knots increases fastly with the degree of the B-spline due to the use of a quasi-interpolation operator to achive an interpolation operator that reproduces the polynomials in the spline space. Then, we propose a direct construction of the spline interpolant in the space

$$V_n := \left\{ \sum_{j\in J} c_j M_n(2\cdot - j) : c_j \in \mathbb{R} \right\}$$

having two useful properties in practice: (a) The fundamental function L_n of the interpolation operator is again a compactly supported function with a small support, but larger than the corresponding one in the construction done in [2] and (b) L_n is symmetric.

As proved in [4], the existence of a solution to

$$\Phi_0 \sum_{j\in\mathbb{Z}} c_{2j} e_{2j} + \Phi_1 \sum_{j\in\mathbb{Z}} c_{2j+1} e_{2j+1} = 1$$

is equivalent to the requirement that the polynomials $\Psi(z)$ and $\Psi(-z)$, where $\Psi(z) := \sum_{j\in\mathbb{Z}} M_n(j) z^j$, have no common zeros in $\mathbb{C}\setminus\{0\}$. Moreover, the equation involving the Laurent polynomials Φ_0 and Φ_1 is equivalent to the identity

$$d(z)\Psi(z) + d(-z)\Psi(-z) = 2,$$

where $d(x) := \sum_{j\in\mathbb{Z}} d_j z^j$. In other words, any sequence $(d_j)_{j\in\mathbb{Z}}$ satistying the equation above provides the fundamental function $\widetilde{L}_n$ of an interpolation operator $\widetilde{\mathcal{L}}_n$. The main problem is to determine a sequence $(d_j)_{j\in\mathbb{Z}}$ such that $\widetilde{\mathcal{L}}_n$ reproduces the polynomials in

 ISBN: 978-84-616-9216-3

$\mathbb{P}_{n-1}$. Since the B-spline M_n is a continuous function of compact support such that its Fourier transform $\widehat{M}_n$ satisfies the conditions $\widehat{M}_n(0) = 1$ and $\widehat{M}_n(\pi) \neq 0$, and

$$\widehat{M}_n^{(\beta)}(2\pi\alpha) = 0,\ \alpha \in \mathbb{Z} \setminus \{0\},\ 0 \leq \beta \leq n-1,$$

then (see [4]) the operator $\widetilde{L}_n$ is exact on $\mathbb{P}_{n-1}$ if and only if

$$d(1) = 2, \quad d^{(\beta)}(-1) = 0,\ 0 \leq \beta \leq n-1.$$

This result is used to derive QDMs for B-splines of low degree. For example, when $n = 5$,

$$\Psi(z) = \frac{1}{384}z^{-2} + \frac{19}{96}z^{-1} + \frac{115}{192} + \frac{19}{96}z + \frac{1}{384}z^2$$

and we look for a Laurent polynomial $d(z) = \sum_{j=-7}^{7} d_j z^j$, $d_j = d_{-j}$, satisfying all the conditions above.

The following solution is obtained:

$$d_0 = \frac{440801}{350208},\ d_1 = \frac{22140839}{35487744},\ d_2 = -\frac{231233}{1400832},\ d_3 = -\frac{5020187}{35487744},$$
$$d_4 = \frac{29215}{700416},\ d_5 = \frac{1861277}{106463232},\ d_6 = -\frac{8383}{1400832},\ d_7 = \frac{8383}{106463232}.$$

Therefore,

$$L_5 = \sum_{j=-7}^{7} d_j M_5(2\cdot - j)$$

and

$$\widetilde{\mathcal{L}}_5 f = \sum_{i\in\mathbb{Z}} f(i)\, L_5(\cdot - i),$$

from which the following DQM results:

$$f'(i) \simeq -\frac{8383}{47316992}(f(i+4) - f(i-4)) + \frac{3767029}{212926464}(f(i+3) - f(i-3))$$
$$- \frac{32434753}{212926464}(f(i+2) - f(i-2)) + \frac{160182545}{212926464}(f(i+1) - f(i-1)).$$

It is a formula exact on $\mathbb{P}_4$. The corresponding one constructed in [2] from $\mathcal{L}_5$ also uses the values of f at $i = -5, 5$. Moreover, it can be proved that

$$\left| f'(i) - \left(\widetilde{\mathcal{L}}_5 f\right)'(i) \right| \leq 0.017326 \left\| f^{(5)} \right\|_{\infty,[i-4,i+4]}$$

and

$$\left| f'(i) - (\mathcal{L}_5 f)'(i) \right| \leq 0.018503 \left\| f^{(5)} \right\|_{\infty,[i-4,i+4]}$$

where $\|g\|_{\infty,I} := \max_{x\in I} |g(x)|$, and so the new construction provides also a better result with respect the constant in the error estimates.

 ISBN: 978-84-616-9216-3

References

[1] D. Barrera and F. Ibáñez, *Compactly supported fundamental functions for spline-based differential quadrature*, in III European Conference on Computational Mechanics Solids, Structures and Coupled Problems in Engineering, C.A. Mota Soares et. al. (eds.), Lisbon, Portugal, 5–8 June 2006.

[2] D. Barrera, P. González, F. Ibáñez and M. J. Ibáñez, J. Comput. Appl. Math. (2014), http://dx.doi.org/10.1016/j.cam.2014.02.006

[3] P.L. Butzer, M. Schmidt, E.L. Stark and L. Vogt, *Central factorial numbers; their main properties and some applications*, Numer. Funct. Anal. and Optimiz. **10** (5&6) (1989) 419–488.

[4] C. A. Micchelli, *Banded matrices with banded inverses*, J. Comput. Appl. Math. **41** (1992) 281–300.

[5] P. Sablonnière, *Spline quasi-interpolants on uniform partitions*, preprint IRMAR 00-38, 2000 (in Spanish).

[6] I. J. Schoenberg, *Cadinal spline interpolation*, SIAM, 1973.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Resolution of parabolic and hyperbolic PDEs using interpolating transient PS-splines

D. Barrera[1], P. González[1], A. Palomares[1] and M. Pasadas[1]

[1] *Department of Applied Mathematics, University of Granada*

emails: dbarrera@ugr.es, prodelas@ugr.es, anpalom@ugr.es, mpasadas@ugr.es

Abstract

In this work we present a procedure to obtain a transient C^1 surface on a polygonal domain Ω which interpolates certain data set and solves numerically a parabolic or hyperbolic second or fourth-order PDE problem considered in this domain. For each instant time considered, the approximation space is in the C^1-quadratic spline space, constructed from an α-triangulation of Ω and its associated Powell-Sabin subtriangulation. That is, using the well known method of lines, we can obtain a system of EDOs in time that, once properly discretized and approximated, permits us to obtain a quite smooth numerical approximation of the original PDE.

Key words: transient PDEs, interpolating PS-splines, Powell-Sabin FE

1 Introduction

In this work we present a procedure to obtain a C^1-surface on a polygonal domain $\Omega \subset \mathbb{R}^2$, depending on time, that also solves the corresponding Galerkin variational formulation of a transient PDE problem up to fourth-order. The approximation space is that of C^1-quadratic splines constructed from the Powell-Sabin subtriangulation associated with an α-triangulation of Ω. We will also use the appropriate interpolation conditions, both on the interior of the domain or over some points on the boundary, in order to take into account the initial and the boundary conditions of each of these problems.

2 Notation and preliminaries

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain (an open polygonal connected set) and let us consider the Sobolev space $H^2(\Omega)$, whose elements are (classes of) functions u defined on Ω such that their partial derivatives (in the distribution sense) $\partial^\gamma u \equiv \frac{\partial^{|\gamma|} u}{\partial x^{\gamma_1} \partial y^{\gamma_2}}$ up to second order ($|\gamma| := \gamma_1 + \gamma_2 \leq 2$) belong to $L^2(\Omega)$.

We will denote $\langle \cdot \rangle$ the usual Euclidean norm and $\langle \cdot\ ,\ \cdot \rangle$ the Euclidean inner product in $\mathbb{R}^2$ and we consider in $H^2(\Omega)$ the usual inner semi-products defined as

$$(u,v)_m := \sum_{|\gamma|=m} \iint_\Omega \partial^\gamma u \cdot \partial^\gamma v, \quad m = 0,1,2;$$

the seminorms

$$|u|_m := (u,u)_m^{1/2} = \left(\sum_{|\gamma|=m} \iint_\Omega (\partial^\gamma u)^2 \right)^{1/2}, \quad m = 0,1,2;$$

and the norm

$$\|u\| = \left(\sum_{m=0}^{2} |u|_m^2 \right)^{1/2} = \left(\sum_{|\gamma|\leq 2} \iint_\Omega (\partial^\gamma u)^2 \right)^{1/2}.$$

Given $\alpha \geq 1$, let $\mathcal{T}$ be an α-triangulation of $\overline{\Omega}$, i. e., a triangulation that satisfies the condition $1 \leq R_T/2r_T \leq \alpha$ for all closed triangles $T \in \mathcal{T}$, R_T and r_T being respectively the radii of the circumscribed and inscribed circles of T, (see e. g. [10]), and let $V_\mathcal{T}$ be the set of all the nodes of $\mathcal{T}$.

We will consider the associated Powell-Sabin subtriangulation $\mathcal{T}'$ of $\mathcal{T}$ (see e. g. [8]), which is obtained by joining the centre Ω_T of the inscribed circle of each interior triangle $T \in \mathcal{T}$ to the vertices of T and to the centres $\Omega_{T'}$ of the inscribed circles of the neighbouring triangles $T' \in \mathcal{T}$. When T has a side lying on the boundary of Ω, the point Ω_T is joined to the mid-point of this side, to the vertices of T and to the centres $\Omega_{T'}$ of the inscribed circles of the neighbouring triangles $T' \in \mathcal{T}$. Hence, all the micro-triangles inside any $T \in \mathcal{T}$ have the incenter of T as a common vertex.

It is well known ([9]) that given the values of a sufficiently smooth function f (defined on $\overline{\Omega}$) and all its first partial derivatives at all the points of $V_\mathcal{T}$, there exists a unique S in

$$\mathcal{S}_2^1(\Omega, \mathcal{T}') = \left\{ S \in \mathcal{C}^1(\Omega) : S|_{T'} \in \mathbb{P}_2(T') \quad \forall T' \in \mathcal{T}' \right\},$$

where $\mathbb{P}_2(T')$ stands for the space of polynomials of total degree at most two over T', such that the values of S and all its first partial derivatives coincide with those of f at all the points of $V_\mathcal{T}$.

 ISBN: 978-84-616-9216-3

3 Resolution of parabolic and hyperbolic second or fourth-order boundary value problems

In this section we formulate and solve numerically both parabolic or hyperbolic second and fourth-order boundary-value problems. These type of transient PDEs arise in a great variety of physical and engineering situations: electric, potential, fluids and elasticity theory (within the study of thin plates), among many others.

3.1 Formulation of the considered problems

Problem 3.1 *Consider the following boundary-value problems in a bounded polygonal domain $\Omega \subset \mathbb{R}^2$, with boundary $\Gamma \equiv \partial\Omega$*

$$\begin{cases} -\frac{\partial u}{\partial t} + \tau_2\,\Delta^2 u - \tau_1\,\Delta u = f,\ t>0 & \text{in} \quad \Omega \\ u\,(t,\cdot) = \phi\,(t,\cdot)\,, \tau_2\,\dfrac{\partial u}{\partial n}\,(t,\cdot) = \tau_2\,\psi\,(t,\cdot)\,, \quad t \geq 0 & \text{on} \quad \Gamma \\ u\,(0,\cdot) = u_0\,(\cdot)\,, & \text{on} \quad \Omega \end{cases} \tag{1}$$

$$\begin{cases} -\frac{\partial^2 u}{\partial t^2} + \tau_2\,\Delta^2 u - \tau_1\,\Delta u = f,\ t>0 & \text{in} \quad \Omega \\ u\,(t,\cdot) = \phi\,(t,\cdot)\,, \tau_2\,\dfrac{\partial u}{\partial n}\,(t,\cdot) = \tau_2\,\psi\,(t,\cdot)\,, \quad t \geq 0 & \text{on} \quad \Gamma \\ u\,(0,\cdot) = u_0\,(\cdot)\,,\ \frac{\partial}{\partial t}u\,(0,\cdot) = u_1\,(\cdot)\,, & \text{on} \quad \Omega \end{cases} \tag{2}$$

with sufficiently regular functions f, ϕ, ψ, u_0, u_1 (see for example [4]) and τ_1, $\tau_2 \geq 0$ are real non negative numbers not vanishing simultaneously.

For solving numerically any of these two types of transient PDE problems in a finite temporal interval $[0,T] \subset \mathbb{R}$ (with $T>0$), we will apply a general Galerkin procedure to their corresponding variational formulation: consider $v \in H_0^2(\Omega)$ and multiply both sides of the corresponding PDE, denoting $\partial_t^{(1)}(\cdot) \equiv \frac{\partial(\cdot)}{\partial t}$ or $\partial_t^{(2)}(\cdot) \equiv \frac{\partial^2(\cdot)}{\partial t^2}$, depending on which Problem (1) or (2) we are considering. Integrating now any of them in the domain $\Omega \subset \mathbb{R}^2$ (with $l=1,2$, depending on the problem considered)

$$\iint_\Omega \left(-\partial_t^{(l)} u + \tau_2 \Delta^2 u - \tau_1\,\Delta u\right)\,v = \iint_\Omega f\,(t,\cdot)\,v\,.$$

It will suffice to apply the appropriate Green formulae to obtain

$$\begin{aligned} \varphi(v)\,(t) &:= \iint_\Omega f\,(t,\cdot)\,v \\ &= -\partial_t^{(l)} \iint_\Omega u\,(t,\cdot)\,\,v + \mathcal{A}(u\,(t,\cdot)\,,v) \end{aligned}$$

 ISBN: 978-84-616-9216-3

where they appear the following bilinear form

$$\begin{aligned}&\mathcal{A}(u\,(t,\cdot)\,,v) := \iint_{\Omega}(\tau_2\,\Delta u\,(t,\cdot)\;\Delta v\;+\tau_1\;\langle\nabla u\,(t,\cdot)\,,\nabla v\rangle)\\&\equiv \tau_1(u\,(t,\cdot)\,,v)_1+\tau_2(\Delta u\,(t,\cdot)\,,\Delta v)_0\end{aligned}\tag{3}$$

and the linear one (depending on the function f)

$$\varphi\,(v)\,(t) := \iint_{\Omega} f\;(t,\cdot)\;v\;\equiv (f\,(t,\cdot)\,,v)_0.\tag{4}$$

So that we can consider now the corresponding variational formulation of the problems (1) or (2), as follows:

Problem 3.2 *To find* $u\,(t,\cdot)\in\mathcal{V}_t$ *such that*

$$u(0,\cdot)=u_0(\cdot),\quad (l-1)\frac{\partial u}{\partial t}(0,\cdot)=(l-1)u_1(\cdot),\tag{5}$$

and for each $t\in\,]0,T[$ *(*$l=1$ *or* 2*, depending on the problem (1) or (2) considered)*

$$\partial_t^{(l)}\iint_{\Omega}u\,(t,\cdot)\;v=\mathcal{A}(u\,(t,\cdot)\,,v)-\varphi\,(v)\,(t)\,,\qquad\forall v\in H_0^2(\Omega)\tag{6}$$

where

$$\mathcal{V}_t\equiv\left\{u\,(t,\cdot)\in H^2(\Omega)\;:\;u\,(t,\cdot)_{|\Gamma}=\phi\,(t,\cdot)\;,\;\frac{\partial u}{\partial n}\,(t,\cdot)_{|\Gamma}=\psi\,(t,\cdot)\right\}.\tag{7}$$

3.2 General settings

Now, let us suppose that we have an appropriate α-triangulation $\mathcal{T}$ of Ω, with its associated Powell-Sabin subtriangulation $\mathcal{T}'$, $D_1=(a_{i_1})_{i_1=1}^{k_1}$ a set of points of $\mathcal{T}$ in $\partial\Omega$ containing all the nodes on this boundary and $D_2=(b_{i_2})_{i_2=1}^{k_2}$ a set of points on $\partial\Omega$ containing all the boundary nodes of $\mathcal{T}'$ that are not geometrical corners of $\partial\Omega$ but in such a way that there is at least one point of D_2 in the interior of every segment of the subtriangulation $\mathcal{T}'$ lying on the $\partial\Omega$, so that we can ensure that every element in

$$\mathcal{H}_0\equiv\left\{\begin{array}{c}v\in\mathbb{S}_2^1(\Omega,\mathcal{T}')\;:\;v\,(a_{i_1})=0=\frac{\partial}{\partial n}v\,(b_{i_2})\,,\\ i_1=1,\ldots,k_1,\;i_2=1,\ldots,k_2\end{array}\right\}$$

will also be in $H_0^2\;(\Omega)$, see [6] for the details.

In this setting, consider then the so called *method of lines* for the variational formulation (5)–(6)–(7) of any of the Problems (1) or (2), with

$$\beta_{i_1}^{(1)}\,(t)=\phi(t,a_{i_1}),\;\forall t\in[0,T]\,,\quad i_1=1,\ldots,k_1,\tag{8}$$

and

$$\beta_{i_2}^{(2)}\,(t)=\psi(t,b_{i_2}),\;\forall t\in[0,T]\,,\quad i_2=1,\ldots,k_2.\tag{9}$$

 ISBN: 978-84-616-9216-3

3.3 Computations

Let $N = dim(\mathbb{S}_2^1(\Omega, \mathcal{T}')) = 3n$ (where n is the number of total nodes of the triangulation $\mathcal{T}$) and we consider the usual Hermite basis $\{B_i\}_{i=1}^N$ of $\mathbb{S}_2^1(\Omega, \mathcal{T}')$. Suppose also that $\{B_{i_1}\}_{i_1=1}^{k_1}$ are the basis functions associated with the degree of freedom $v \mapsto v(a_{i_1})$, for $i_1 = 1, \ldots, k_1$ and $\{B_{i_2+k_1}\}_{i_2=1}^{k_2}$ the basis functions associated with the data $v \mapsto \dfrac{\partial v}{\partial n}(b_{i_2})$, for $i_2 = 1, \ldots, k_2$.

It is easy to check that if we denote $k := k_1 + k_2$ then $\{B_{i+k}\}_{i=1}^{N-k}$ is a basis of $\mathcal{H}_0$, and hence, we could express

$$\widetilde{u}\,(t,\cdot) = \sum_{i_1=1}^{k_1} \beta_{i_1}^{(1)}\,(t)\,B_{i_1}\,(\cdot) + \sum_{i_2=1}^{k_2} \beta_{i_2}^{(2)}\,(t)\,B_{i_2+k_1}\,(\cdot) + \sum_{i=1}^{N-k} c_i\,(t)\,B_{i+k}\,(\cdot)\,. \tag{10}$$

So that, if we take $v = B_{j+k}$, for some $j \in \{1, \ldots, N-k\}$ (with $l = 1$ or 2 depending on the problem considered), we have from (6)

$$\partial_t^{(l)} \iint_\Omega \widetilde{u}\,(t,\cdot)\ B_{j+k} = \mathcal{A}(\widetilde{u}\,(t,\cdot)\,, B_{j+k}) - \varphi\,(B_{j+k})\,(t)\,.$$

So, developing a little bit more these expressions we get that the vector $\mathbf{c}\,(t) = (c_i\,(t))_{i=1}^{N-k}$ is the solution of the linear system of ordinary differential equations

$$B\,\partial_t^{(l)}\mathbf{c}\,(t) - A\,\mathbf{c}\,(t) = \mathbf{b}\,(t)\,, \tag{11}$$

with the matrices $A, B \in \mathbb{R}^{N-k,N-k}$

$$A = (\mathcal{A}\,(B_{i+k},\ B_{j+k}))_{1\le i,j\le N-k}\,; \quad B = \big((B_{i+k},\ B_{j+k})_0\big)_{1\le i,j\le N-k}$$

and the components of the vector $\mathbf{b}\,(t) \equiv (b_j\,(t))_{1\le j\le N-k}^\top \in \mathbb{R}^{N-k}$ are defined by

$$b_j\,(t) = -\left(\sum_{i_1=1}^{k_1} \beta_{i_1}^{(1)}\,(t)\ \mathcal{A}\,(B_{i_1}, B_{j+k}) + \sum_{i_2=1}^{k_2} \beta_{i_2}^{(2)}\,(t)\ \mathcal{A}\,(B_{i_2+k_1}, B_{j+k}) + \varphi(B_j)\,(t)\right)_{j=k+1}^{N}.$$

Clearly, the coefficient matrices A and B are symmetric, banded (this is due to the fact that each B_i has local support), and positive definite.

Concerning the initial conditions associated to this system of ordinary differential equations (11) in both problems (1) and (2), we just remember that evaluating (10) at time $t = 0$ we get, taking into account (8) and (9),

$$\begin{aligned}\widetilde{u}\,(0,\cdot) &= \sum_{i_1=1}^{k_1} \beta_{i_1}^{(1)}\,(0)\,B_{i_1}\,(\cdot) + \sum_{i_2=1}^{k_2} \beta_{i_2}^{(2)}\,(0)\,B_{i_2+k_1}\,(\cdot) + \sum_{i=1}^{N-k} c_i\,(0)\,B_{i+k}\,(\cdot)\\ &= \sum_{i_1=1}^{k_1} \phi(0, a_{i_1})\,B_{i_1}\,(\cdot) + \sum_{i_2=1}^{k_2} \psi(0, b_{i_2})\,B_{i_2+k_1}\,(\cdot) + \sum_{i=1}^{N-k} c_i\,(0)\,B_{i+k}\,(\cdot)\end{aligned}$$

 ISBN: 978-84-616-9216-3

and just in the case of the hyperbolic problem (2)

$$\begin{aligned}\frac{\partial}{\partial t}\widetilde{u}\,(0,\cdot) &= \sum_{i_1=1}^{k_1}\left(\beta_{i_1}^{(1)}\right)'(0)\,B_{i_1}\,(\cdot)+\sum_{i_2=1}^{k_2}\left(\beta_{i_2}^{(2)}\right)'(0)\,B_{i_2+k_1}\,(\cdot)+\sum_{i=1}^{N-k}c_i'\,(0)\,B_{i+k}\,(\cdot)\\ &= \sum_{i_1=1}^{k_1}\frac{\partial}{\partial t}\phi(0,a_{i_1})\,B_{i_1}\,(\cdot)+\sum_{i_2=1}^{k_2}\frac{\partial}{\partial t}\psi(0,b_{i_2})\,B_{i_2+k_1}\,(\cdot)+\sum_{i=1}^{N-k}c_i'\,(0)\,B_{i+k}\,(\cdot)\,.\end{aligned}$$

So, if we want that any or both of these two initial conditions could be well approximated by the so denoted quadratic spline functions $\widetilde{u}_0,\widetilde{u}_1\in\mathcal{S}_2^1(\Omega,\mathcal{T}')$ so that we could write

$$\widetilde{u}\,(0,\cdot)=\widetilde{u}_0\,(\cdot)\simeq u_0\,(\cdot)\,,\quad \frac{\partial}{\partial t}\widetilde{u}\,(0,\cdot)=\widetilde{u}_1\,(\cdot)\simeq u_1\,(\cdot) \tag{12}$$

it will suffice to take the appropriate coefficients $\{c_i\,(0)\}_{i=1}^{N-k}$ and $\{c_i'\,(0)\}_{i=1}^{N-k}$ in order to verify (12) for the unique $\widetilde{u}_0$, $\widetilde{u}_1\in\mathcal{S}_2^1(\Omega,\mathcal{T}')$ interpolating u_0,u_1 in the sense that their values and that of their first partial derivatives coincide in all the nodes of the triangulation $\mathcal{T}$ considered (see [8] or [9]).

Acknowledgements

Work partially supported by Junta de Andalucía (Research group FQM/191) and by the Ministerio de Economía y Competitividad of Spain under grant MTM2011-26468. Third author also acknowledges support from the Research group TEP190.

References

[1] K. Atkinson and W. Han, *Theoretical Numerical Analysis*, Springer, 2nd. edition, 2005.

[2] R. Arcangéli, M.C. López de Silanes and J.J.Torrens, *Multidimensional minimizing splines*, Kluwer Academic Publisher, 2004.

[3] D. Barrera, M. A. Fortes, P. González and M. Pasadas, *Minimal energy surfaces on Powell-Sabin triangulations*, Applied Numerical Mathematics **58** (2008) 635-645.

[4] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Dunod, Paris, 1999. Nouvelle présentation 2005.

[5] J. Duchon, *Interpolation des fonctions de deux variables suivant le principle de la flexion des plaques minces*, RAIRO **10 (12)** (1976) 5–12.

 ISBN: 978-84-616-9216-3

[6] M. A. Fortes, P. González, M.J. Ibañez and M. Pasadas, *Interpolating minimal energy $\mathcal{C}^1$-surfaces on Powell-Sabin triangulations. Applications to the resolution of elliptic problems*, Submitted to Numerical Methods on Partial Differential Equations.

[7] A. Kouibia, M. Pasadas, *Approximation by interpolating variational splines*, J. Comput. Appl. Math. **218** (2008) 342-349.

[8] M. Laghchim-Lahlou and P. Sablonnière, *C^r-finite elements of Powell-Sabin type on the three direction mesh*, Adv. in Comput. Math. **6** (1996) 191–206.

[9] M.J.D. Powell and M.A. Sabin, *Piecewise Quadratic Approximations on Triangles*, ACM Transactions on Mathematical Software **3(4)** (1977) 316–325

[10] P. Sablonnière, *Error bounds for Hermite interpolation by quadratic splines on an α-triangulation*, IMA Journal of Numerical Analysis **7 (4)** (1987) 495–508.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Approximation of Multivariate Functions via Fluctuationlessness Theorem by Using Nested Taylor Decomposition

N. A. Baykara[1] **and Ercan Gürvit**[1]

[1] *Marmara University, Mathematics Department.*

emails: nabaykara@gmail.com, ercangurvit@gmail.com

Abstract

In this paper the scheme to approximate univariate functions via fluctuationlessness theorem applied on the nested remainder term of Taylor decomposition of an analytic function. We extend the general scheme to multivariate functions by using one dimensional Taylor expansion not to the independent variables but to a parameter characterizing directional changes in the function values. Certain remarks are given on the subject.

Key words: Fluctuationlessness theorem, Nested Taylor decomposition, Remainder term, matrix representation of a function, Approximation of functions, Multivariate approximation

1 Introduction

Many papers [1-3] from our working group are written about the possibility of using Taylor series remainder term evaluation via fluctuation free integration in the univariate integration of the functions even for the cases where Taylor polynomials present very poor approximation quality. What we have done there is now considered not only for integration but function approximation. We had already proposed an approach to approximate a univariate function by using Taylor's expansion and utilizing the fluctuation free integration approximation for the explicit expression of Taylor's remainder term. Moreover in a new article we have proposed a new addition to this method by adding a nested Taylor decomposition applied to the integrand of the remainder term. In this work we develop a similar method for approximating functions of many variables. What we produce here is applicable to the multivariate integration even though there seem to exist a lot of geometrical limitations which urge us to develop a more comprehensive algorithm to that end.

2 Multivariate Taylor Decomposition

Let $\mathbf{x} = (x_1, x_2, ..., x_N)$ lie in the ball B with center $\mathbf{a} = (a_1, a_2, ..., a_N)$ and f be a real-valued function defined on the closure $\overline{B}$ having all continuous partial derivatives up to and including $(k+1)$th order at every point. Now Taylor's theorem for a single variable can be applied to $f(\mathbf{v}(t))$ by expanding it about $t = 0$ up to the kth term and evaluating it at $t = 1$ It is clear that in our present work this formulation will only be applied for two variables

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{v}(1)) = f\left(\mathbf{x}^{(0)}\right) &+ \sum_{i=1}^{k} \frac{1}{i!} \left.\frac{d^i f(\mathbf{v}(t))}{dt^i}\right|_{t=0} \\ &+ \int_0^1 dt \frac{(1-t)^k}{k!} \frac{d^{k+1} f(\mathbf{v}(t))}{dt^{k+1}} \end{aligned} \tag{1}$$

To proceed further, the chain rule for several variables is to be applied to the ith derivative of $f(\mathbf{v}(t))$

$$\begin{aligned} \frac{d^i}{dt^i} f(\mathbf{v}(t)) &= \frac{d^i}{dt^i} f\left(\mathbf{x}^{(0)} + t\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) \\ &= \sum_{|\sigma|=i} \begin{pmatrix} \mathrm{i} \\ \sigma \end{pmatrix} \left(\mathbf{x} - \mathbf{x}^{(0)}\right)^{\sigma} \left(D^{\sigma} f\right)\left(\mathbf{x}^{(0)} + t\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) \end{aligned} \tag{2}$$

Corresponding notations can be found in [1]. Now, using the equality given by (1) $f(\mathbf{x})$ can be written as the sum of a kth degree Taylor polynomial $P_k(\mathbf{x})$ and a remainder term $R_k(\mathbf{x})$; namely

$$f(\mathbf{x}) = P_k(\mathbf{x}) + R_k(\mathbf{x}) \tag{3}$$

where

$$P_k(\mathbf{x}) = f\left(\mathbf{x}^{(0)}\right) + \sum_{|\sigma|=1}^{k} \frac{\left(\mathbf{x} - \mathbf{x}^{(0)}\right)^{\sigma}}{\sigma!} \left(D^{\sigma} f\right)\left(\mathbf{x}^{(0)}\right) \tag{4}$$

and

$$\begin{aligned} R_k(\mathbf{x}) = &\sum_{|\sigma|=k+1} \frac{k+1}{\sigma!} \left(\mathbf{x} - \mathbf{x}^{(0)}\right)^{\sigma} \\ &\times \int_0^1 dt (1-t)^k \left(D^{\sigma} f\right)\left(\left(\mathbf{x} - \mathbf{x}^{(0)}\right) t + \mathbf{x}^{(0)}\right) \end{aligned} \tag{5}$$

This remainder term will be modified to a more convenient form so that the Fluctuation-lessness Theorem can be easily applied. For this purpose a weight function is to be defined

$$w_k(t) \equiv (k+1)(1-t)^k, \qquad k = 0, 1, 2, ... \tag{6}$$

 ISBN: 978-84-616-9216-3

This allows (5) to be reformulated as

$$R_k(\mathbf{x}) = \sum_{|\sigma|=k+1} \frac{(\mathbf{x}-\mathbf{x}^{(0)})^\sigma}{\sigma!} \int_0^1 dt w_k(t) \times (D^\sigma f)\left(\left(\mathbf{x}-\mathbf{x}^{(0)}\right)t + \mathbf{x}^{(0)}\right) \tag{7}$$

Letting,

$$\mathcal{I}_{k,\sigma}(\mathbf{x}) = \int_0^1 dt w_k(t) (D^\sigma f)\left(\left(\mathbf{x}-\mathbf{x}^{(0)}\right)t + \mathbf{x}^{(0)}\right) \tag{8}$$

the remainder term becomes

$$R_k(\mathbf{x}) = \sum_{|\sigma|=k+1} \frac{(\mathbf{x}-\mathbf{x}^{(0)})^\sigma}{\sigma!} \mathcal{I}_{k,\sigma}(\mathbf{x}) \tag{9}$$

3 Fluctuationlessness Theorem

The fluctuation free matrix representation approximation is based on a theorem which was conjectured and proven by M. Demiralp. This theorem states that the matrix representation of an algebraic operator which multiplies its argument by a scalar multivariate function, is identical to the image of the independent variables' matrix representations over the same space via the same basis set, under that multivariate function, when the fluctuation terms are ignored. This is in fact the multivariate counterpart of the fluctuationlessness theorem for univariate functions, which was conjectured and proven by the same author.

The details about the theorem for the univariate case were given in previous works of our group. Here we give the extension to the multivariate case in order to emphasize on the generality of the theorem without depending on the multivariance except certain extensions in the mathematical objects, even though we are not going to use this extended form in the analysis here. Let g be a multivariate function defined over a rectangular hyperprism, say $[a_1, b_1] \times \cdots \times [a_n, b_n]$ where n is the number of the independent variables, and analytic throughout its domain. We define $\mathbf{u}(x_1, ..., x_n) = [u_1(x_1), u_2(x_2), ...]^T$ such that $u_i(x_1, ..., x_n)$'s are orthonormal basis functions of the Hilbert space from which the function g is chosen. We can define the algebraic function multiplication operator $\widehat{g}$ whose action on its operand is the multiplication with the value of $g(x_1, ..., x_n)$. We can also define a matrix representation operator, $\widehat{\mathbf{M}}(\widehat{g})$, which maps from the function operator $\widehat{g}$ to an infinite matrix defined as the following inner product matrix

$$\widehat{\mathbf{M}}(\widehat{g}) \equiv \left(\mathbf{u}, \widehat{g}\mathbf{u}^T\right) \tag{10}$$

where the (i, j)-th component of the image matrix is defined as the inner product

$$\int_{\mathcal{V}} d\mathcal{V} w(x_1, ..., x_n) u_i(x_1, ..., x_n) g(x_1, ..., x_n) u_j(x_1, ..., x_n). \tag{11}$$

 ISBN: 978-84-616-9216-3

The arguments being the matrix representation of the variables x_1,...,x_n we can write the above approximation as

$$\widehat{\mathbf{M}}\left(\widehat{g}\right) \approx g\left(\mathbf{X}_1, ..., \mathbf{X}_n\right) \tag{12}$$

where $\mathbf{X}$s are the matrix representations of $\widehat{x}$s. As n goes to infinity the approximation becomes an exact equality.

Even though we never distinguished the composite function matrix representations, we can give the following equation for the composite function say $f\left(g\left(x\right)\right)$ derived from the univariate functions $f(x)$ and $g(x)$ at the fluctuationlessness limit

$$\widehat{\mathbf{M}}\left(\widehat{f}\left(\widehat{g}\right)\right) \approx f\left(\widehat{\mathbf{M}}\left(\widehat{g}\right)\right) \tag{13}$$

which may be used instead of the expression in terms of universal matrix (the matrix representation of the independent variable)

$$\widehat{\mathbf{M}}\left(\widehat{f}\left(\widehat{g}\right)\right) \approx f\left(g\left(\mathbf{X}\right)\right). \tag{14}$$

(13) may give better approximation results since (14) uses the two nested fluctuationlessness application while (13) skips the inner one.

4 Nested Taylor Formulation

Consider now the Taylor decomposition of a function $f(x)$ as the sum of an n_1-th Taylor Polynomial around a point x_1 and a remainder term expressed in integral form. All throughout this work the necessary continuity conditions will be assumed to hold [4,5].

$$f\left(x\right) = \sum_{j=0}^{n_1} \frac{f^{(j)}\left(x_1\right)}{j!}\left(x - x_1\right)^j + \frac{1}{n_1!}\int_{x_1}^{x} dt_1 \left(x - t_1\right)^{n_1} f^{(n_1+1)}\left(t_1\right), \qquad n_1 = 0, 1, \ldots \tag{15}$$

The next step is the second Taylor decomposition of the function appearing in the integral above. The function $f^{(n_1+1)}\left(t_1\right)$ is expanded around yet another point (call it x_2) and another remainder term expressed once again in integral form.

$$f^{(n_1+1)}\left(t_1\right) = \sum_{j=0}^{n_2} \frac{f^{(n_1+1+j)}\left(x_2\right)}{j!}\left(t_1 - x_2\right)^j + \frac{1}{n_2!}\int_{x_2}^{t_1} dt_2 \left(t_1 - t_2\right)^{n_2} f^{(n_1+n_2+2)}\left(t_2\right) \tag{16}$$

These will yield the following expansion for $f(x)$

$$\begin{aligned} f\left(x\right) &= \sum_{j=0}^{n_1} \frac{f^{(j)}\left(x_1\right)}{j!}\left(x - x_1\right)^j + \frac{1}{n_1!}\int_{x_1}^{x} dt_1 \left(x - t_1\right)^{n_1} \sum_{j=0}^{n_2} \frac{f^{(n_1+1+j)}\left(x_2\right)}{j!}\left(t_1 - x_2\right)^j \\ &+ \frac{1}{n_1! n_2!}\int_{x_1}^{x} dt_1 \int_{x_2}^{t_1} dt_2 \left(x - t_1\right)^{n_1}\left(t_1 - t_2\right)^{n_2} f^{(n_1+n_2+2)}\left(t_2\right) \end{aligned} \tag{17}$$

 ISBN: 978-84-616-9216-3

Now let us proceed with the polynomial parts of the Taylor expansions. We make following definitions.

$$P_{n_1}^{(1)}(x) \equiv \sum_{j=0}^{n_1} \frac{f^{(j)}(x_1)}{j!}(x-x_1)^j \tag{18}$$

$$P_{n_1+n_2+1}^{(2)}(x) \equiv \sum_{j=0}^{n_2} \frac{f^{(n_1+1+j)}(x_2)}{n_1!j!} \int_{x_1}^{x} dt_1 (x-t_1)^{n_1}(t_1-x_2)^j \tag{19}$$

we will now deal with the integral appearing in the expression of $P^{(2)}$. To this end a change of variable can be done by replacing t_1 with t_1+x_2. So the following turns out to be true

$$\int_{x_1}^{x} dt_1 (x-t_1)^{n_1}(t_1-x_2)^j = \int_{x_1-x_2}^{x-x_2} dt_1 (x-x_2-t_1)^{n_1} t_1^j \tag{20}$$

Now (20) can be expressed as the sum of two integrals

$$\int_{0}^{x-x_2} dt_1 (x-x_2-t_1)^{n_1} t_1^j + \int_{x_1-x_2}^{0} dt_1 (x-x_2-t_1)^{n_1} t_1^j \tag{21}$$

Starting with the first of these integrals we can make another change of variable by replacing t_1 with $(x-x_2)\,t_1$ to obtain

$$\begin{aligned}\int_{0}^{x-x_2} dt_1 (x-x_2-t_1)^{n_1} t_1^j &= (x-x_2)^{n_1+j+1}\int_0^1 dt_1 (1-t_1)^{n_1} t_1^j \\ &= (x-x_2)^{n_1+j+1}\,\beta(n_1+1,j+1) \\ &= (x-x_2)^{n_1+j+1}\frac{\Gamma(n_1+1)\Gamma(j+1)}{\Gamma(n_1+j+2)} = \frac{n_1!j!}{(n_1+j+1)!}(x-x_2)^{n_1+j+1}\end{aligned} \tag{22}$$

Now, for the second integral

$$\begin{aligned}\int_{x_1-x_2}^{0} dt_1 (x-x_2-t_1)^{n_1} t_1^j &= -\int_0^{x_1-x_2} dt_1 (x-x_2-t_1)^{n_1} t_1^j \\ &= -(x_1-x_2)^{j+1}\int_0^1 dt_1 (x-x_2-(x_1-x_2)\,t_1)^{n_1} t_1^j \\ &= -(x_1-x_2)^{j+1}(x-x_2)^{n_1}\int_0^1 dt_1 \left(1-\frac{x_1-x_2}{x-x_2}t_1\right)^{n_1} t_1^j\end{aligned} \tag{23}$$

The integral appearing in the last form of (9) can be expressed as

$$\int_0^1 dt_1 \left(1-\frac{x_1-x_2}{x-x_2}t_1\right)^{n_1} t_1^j = \frac{j!}{(j+1)!}\,{}_2F_1\left(-n_1, j+1; j+2; \frac{x_1-x_2}{x-x_2}\right) \tag{24}$$

where ${}_2F_1$ is for the Gauss Hypergeometric Function. And we can express the remainder term as

$$R^{(2)}_{n_1,n_2} \equiv \frac{1}{n_1!n_2!}\int_{x_1}^{x} dt_1 \int_{x_2}^{t_1} dt_2\,(x-t_1)^{n_1}\,(t_1-t_2)^{n_2}\,f^{(n_1+n_2+2)}\,(t_2) \tag{25}$$

This can be decomposed to certain univariate integrals each of which can be approximately evaluated by using the fluctuationless theorem.

5 Concluding Remarks

This nested Taylor formulation can easily be applied to the multivariable form described over here, the argument being a vector of two variables. By fetching through a nested decomposition the main purpose is to be able to obtain a better approximation compared to a regular Fluctuationlessness approximation of multivariate functions.

Acknowledgements

Both authors are grateful to Prof. Metin Demiralp for his highly valuable contributions and discussions.

References

[1] Ercan Gürvit, N.A. Baykara and Metin Demiralp, Evaluation of Multivariate Integrals via Fluctuationlessness Theorem and Taylor's Remainder, AIP Conf. Proc. 1148, pp. 128-132 (2009).

[2] N.A. Baykara, E. Gürvit, M. Demiralp,The fluctuationlessness approach to the numerical integration of functions with a single variable by integrating Taylor expansion with explicit remainder term, J. Math. Chem. 49, pp. 393-406 (2011), DOI: 10.1007/s10910-010-9748-5

[3] M. Demiralp, A Fluctuation Expansion Method for the Evaluation of a Function's Expectation Value, Int. Conf. on Numer. Anal. and Appl.Math., Wiley, Rhodes, Greece, Sept. 16 – 20, 2005, pp. 711-714

[4] Ercan Gürvit, N. A. Baykara,"Nested Taylor Decomposition of Univariate Functions under Fluctuationlessness Approximation", AIP Proceedings, ICCMSE'2014, Accepted(In Print)

[5] N. A. Baykara,Ercan Gürvit, "Numerical Integration Based on Nested Taylor Decomposition of Univariate Functions under Fluctuationlessness Approximation", AIP Proceedings, ICCMSE'2014, Accepted(In Print)

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Sensitivity analysis of a linear unbranched chemical process with n steps

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, EPI Gión, Spain*

emails: bayon@uniovi.es, jaurelio@uniovi.es, mruiz@uniovi.es, pedrosr@uniovi.es, ctasis@uniovi.es

Abstract

In this paper we present a quasi-analytical method to calculate the optimal enzyme concentrations in a chemical process by considering the minimization of the operation time. The resulting constrained optimal control problem is solved using Pontryagin's Minimum Principle. Our method allows us, first, to obtain the generalized solution of a n-step system with an unbranched scheme and bilinear kinetic models and with non-equal catalytic efficiencies of the enzymes. Second, we discuss in detail the sensitivity analysis of these catalytic parameters.

Key words: Optimal Control, Chemical Process, Sensitivity Analysis
MSC 2000: 49J30, 49M05, 92E20, 80A30, 92C40

1 Introduction

Let us consider an unbranched metabolic pathway composed of n irreversible reaction steps converting substrate x_1 into product p. An explicit solution for the simplest case, i.e. $n = 2$, can be found in [1], while for longer pathways, the authors solved the optimization problem numerically. The solution is obtained quasi-analytically in [2], though with the constraint of considering only the case of $n = 3$ with two intermediate compounds. [3] present several theoretical results over qualitative properties of the solution for the general case of n steps. These authors prove that the optimal enzyme concentration profile is of the "bang-bang" type, though they do not present the analytical solution. In a previous paper [4], we extended the theoretical analysis of [1], [2] and [3], presenting the quasi-analytical solution for the more general case of n steps and assuming equal catalytic efficiencies of

the enzymes ($k_i = 1$). We considered the minimization of the transition time in [4]. This transition time is defined by a improper integral running until infinite time. Given that this model is somewhat unreal, in this paper we shall consider a more realistic situation in chemistry or biology. Moreover, we shall substantially extend the theoretical analysis of [4] to consider nonequal catalytic efficiencies k_i.

Sensitivity analysis (SA), on the other hand, investigates the relations between parameters of a model and a property of the outcome. Classically (see, for example, [5]), SA is performed by the partial derivatives of the outcome with respect to its parameters. When a closed-form equation describes the relationship between the independent variables and the dependent variable, this SA is easy to perform. This is precisely the major advantage of our method: it allows us to obtain the partial derivatives of the concentration of the compounds x_i with respect to the catalytic efficiencies of the enzymes k_i.

2 Statement of the Problem

2.1 Model formulation

Let us consider the following unbranched metabolic pathway composed of n irreversible reaction steps converting substrate x_1 into product p:

$$x_1 \overset{u_1}{\to} x_2 \overset{u_2}{\to} x_3 \overset{u_3}{\to} \cdots \to x_{n-1} \overset{u_{n-1}}{\to} x_n \overset{u_n}{\to} p \tag{1}$$

where $x_1(t)$ is the substrate concentration at time t, $p(t)$ the concentration of the final product at time t, $x_i(t)$ $(i = 2, \ldots, n)$ the concentration of the intermediate compounds at time t, and $u_i(t)$ $(i = 1, \ldots, n)$ the concentration at time t of the enzyme catalyzing the i-th reaction. The model of the reactions in (1) can then be described by the set of differential equations:

$$\begin{cases} \dot{x}_1 = -k_1 u_1 x_1 & x_1(0) = 1, \quad x_1(t) \geq 0 \\ \dot{x}_2 = k_1 u_1 x_1 - k_2 u_2 x_2 & x_2(0) = 0, \quad x_2(t) \geq 0 \\ \dot{x}_3 = k_2 u_2 x_2 - k_3 u_3 x_3 & x_3(0) = 0, \quad x_3(t) \geq 0 \\ \cdots & \\ \dot{x}_n = k_{n-1} u_{n-1} x_{n-1} - k_n u_n x_n & x_n(0) = 0, \quad x_n(t) \geq 0 \end{cases} \tag{2}$$

In a previous paper [4], we assumed equal catalytic efficiencies of the enzymes ($k_i = 1$). In this paper, we shall substantially generalize the study to consider nonequal catalytic efficiencies.

2.2 Objective function

Our goal is to convert substrate x_1 into product p as fast as possible and several cost functions may be considered. The *transition time*, τ (defined in [6]), is used in [1], [2] and

 ISBN: 978-84-616-9216-3

[4]. This transition time is defined by a time integral running until infinite time:

$$\min_{u_1,...u_n} \tau = \min_{u_1,...u_n} \int_0^\infty \frac{1}{x_1(0)}(x_1(0) - p(t))dt \tag{3}$$

In this paper, we shall consider a more realistic situation in biology where the product $p(t)$ need not be fully synthesized, but rather synthesized to a defined concentration. We therefore minimize the *operation time* (to distinguish it from the transition time) defined by specifying the final product concentration, e.g. $p(t_f) = 0.9$, with t_f as the final time. The objective function of the optimization problem may thus be defined as:

$$\tau_{90} = \min_{u_1,...u_n} t_f = \min_{u_1,...u_n} \int_0^{t_f} dt \tag{4}$$

3 Optimal Solution

In this section, we present the solution to the optimal control problem (OCP) defined in the previous section:

$$\min_{\mathbf{u}(t)} \int_0^{t_f} F\left(t, \mathbf{x}(t), \mathbf{u}(t)\right) dt \tag{5}$$

subject to satisfying:

$$\dot{\mathbf{x}}(t) = f\left(t, \mathbf{x}(t), \mathbf{u}(t)\right) \tag{6}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{7}$$

$$\mathbf{u}(t) \in \Omega,\ 0 \le t \le t_f \tag{8}$$

where $F \equiv 1$ is the objective function, $\mathbf{x} = (x_1(t), ..., x_n(t)) \in \mathbb{R}^n$ is the state vector, with initial conditions $\mathbf{x}_0$, $\mathbf{u} \in \mathbb{R}^n$ is the control vector, Ω denotes the set of admissible control values and t is the operating time, which starts from 0 and ends at t_f (value to minimize). The state variables must satisfy the state equation (6) with given initial conditions. In this statement, we consider the final state to be free. Let H be the Hamiltonian function associated with the problem

$$H(t, \mathbf{x}, \mathbf{u}, \lambda) = F\left(t, \mathbf{x}, \mathbf{u}\right) + \lambda \cdot f\left(t, \mathbf{x}, \mathbf{u}\right) \tag{9}$$

where $\lambda = (\lambda_1(t), ..., \lambda_n(t)) \in \mathbb{R}^n$ is called the *costate vector*. The classical approach involves the use of Pontryagin's Minimum Principle [7], which results in a two-point boundary value problem (TPBVP). In order for $\mathbf{u} \in \Omega$ to be optimal, a nontrivial function λ must necessarily exist, such that for almost every $t \in [0, t_f]$:

$$\dot{\mathbf{x}} = H_\lambda;\ \mathbf{x}(0) = \mathbf{x}_0 \tag{10}$$

$$\dot{\lambda} = -H_{\mathbf{x}};\ \lambda(t_f) = \mathbf{0} \tag{11}$$

$$\min_{\mathbf{u}\in\Omega} H(t, \mathbf{x}, \mathbf{u}, \lambda) \tag{12}$$

 ISBN: 978-84-616-9216-3

We now present the solution to the optimal control problem defined above using Pontryagin's Minimum Principle [7]. The fundamental result to obtain may be summarized as follows:

Theorem 1. *There exists a set of switching times* $\{t_1, t_2, ..., t_{n-1}\}$, *(with* $0 < t_i < t_j$, *for* $i < j$*) which partition the optimization interval as:*

$$[0, t_1) \cup [t_1, t_2) \cup \cdots \cup [t_{n-2}, t_{n-1}) \cup [t_{n-1}, t_f] \tag{13}$$

such that the optimal profile of the i*-th enzyme is of the bang-bang type and satisfies:*

$$u_i^*(t) = \begin{cases} 1 & for \quad t \in [t_{i-1}, t_i) \\ 0 & for \quad t \notin [t_{i-1}, t_i) \end{cases} ; \; i = 1, \ldots, n \tag{14}$$

with $t_0 = 0$ and $t_n = t_f$. *In each interval* $[t_{i-1}, t_i]$, $i = 1, \ldots, n$, *the optimal metabolite concentration is given by:*

$$x_1(t) = \begin{cases} e^{-k_1 t} & i = 1 \\ e^{-k_1 t_1} & i > 1 \end{cases} \tag{15}$$

$$x_j(t) = \begin{cases} \prod_{h=1}^{j-1} (1 - e^{-k_h(t_h - t_{h-1})}) \cdot e^{-k_j(t_j - t_{j-1})} & j = 2, \ldots, i-1 \\ \prod_{h=1}^{j-1} (1 - e^{-k_h(t_h - t_{h-1})}) \cdot e^{-k_j(t - t_{i-1})} & j = i \\ \prod_{h=1}^{i-1} (1 - e^{-k_h(t_h - t_{h-1})}) \cdot (1 - e^{-k_i(t - t_{i-1})}) & j = i+1 \\ 0 & j = i+2, \ldots, n \end{cases} \tag{16}$$

We have thus solved the problem quasi-analytically. The optimal solution has been obtained analytically for all the intervals $[0, t_1) \cup [t_1, t_2) \cup \cdots \cup [t_{n-1}, t_f]$. The calculation of the switching times $t_1, t_2, \ldots, t_{n-1}$ and the value of t_f is the only one that is not carried out analytically or exactly.

4 Examples

Using the results presented in the previous section, we developed a program using the Mathematica package that allows us to obtain the optimal solution.

4.1 Example 1: Optimal solution

Let us consider the following values for the nonequal catalytic efficiencies k_i:

$$k_1 = 10; \; k_2 = 10; \; k_3 = 9; \; k_4 = 9; \; k_5 = 8; \; k_6 = 7; \; k_7 = 5; \; k_8 = 3; \; k_9 = 12 \tag{17}$$

In Table I, we present the optimal solution for the cases $n = 3, \ldots, 9$. Let us see the switching times t_i $(i = 1, \ldots, n)$, and the operation time $\tau = t_n$. Remember that u_i is given

 ISBN: 978-84-616-9216-3

by 1 in all the intervals (when it is active). Moreover, the substrate concentration, x_1, the concentrations of the intermediate compounds, $x_2, \ldots, x_n$, and the concentration of the final product, p, are immediately obtained in any interval using the formulas presented in Theorem 1. Figure 1 shows the optimal solution for the case $n = 9$.

Table I. Switching times and operation time of the optimal solution.

n	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
3	0.3401	0.6803	1.0469	-	-	-	-	-	-
4	0.3702	0.7404	1.1404	1.5404	-	-	-	-	-
5	0.3958	0.7917	1.2201	1.6485	2.1160	-	-	-	-
6	0.4188	0.8376	1.2915	1.7453	2.241	2.7898	-	-	-
7	0.4440	0.8880	1.3698	1.8516	2.3792	2.9633	3.7150	-	-
8	0.4755	0.9510	1.4677	1.9845	2.5512	3.1801	3.9942	5.1845	-
9	0.4821	0.9642	1.4883	2.0124	2.5874	3.2257	4.0529	5.2648	5.6817

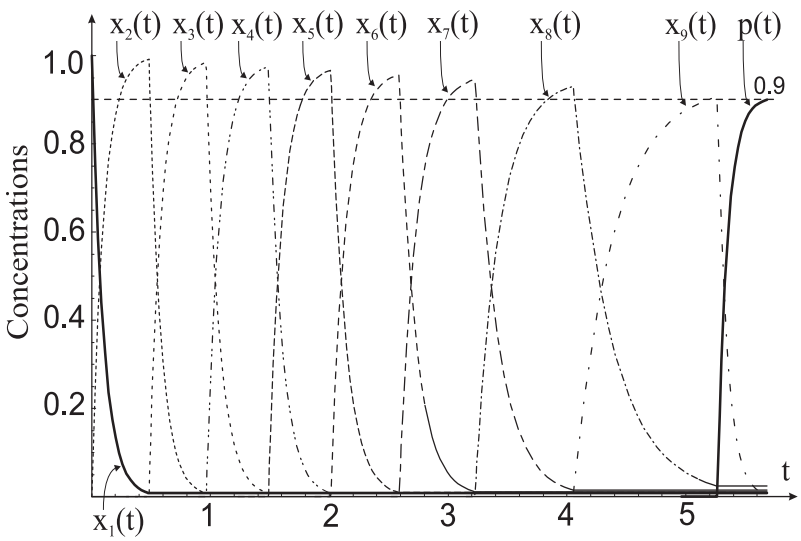


Figure 1. Metabolite and product profile. Case $n = 9$.

4.2 Example 2: Differential SA

Sensitivity analysis (SA) investigates the effect of parameter change on the solution of mathematical models, with more than a dozen SA techniques having been reported ([5]). *Differential SA* will be employed in the present paper. In this case, the sensitivity coefficient, ϕ_i, for a particular independent variable can be calculated from the partial derivative of the dependent variable with respect to the independent variable. When an explicit algebraic equation describes the relationship, the differential SA is easy to perform.

Let us now see how the Differential SA of our problem can be performed immediately, employing analytic formulas to do so (16). The sensitivity coefficient, ϕ_{ij}, defined from the partial derivative of the dependent variable x_i $(i = 1, \ldots, n)$ with respect to k_j $(i = 1, \ldots, i)$:

$$\phi_{ij} = \frac{dx_i}{k_j} \tag{18}$$

was calculated using the Mathematica package. A summary of the results is shown in Figure 2.

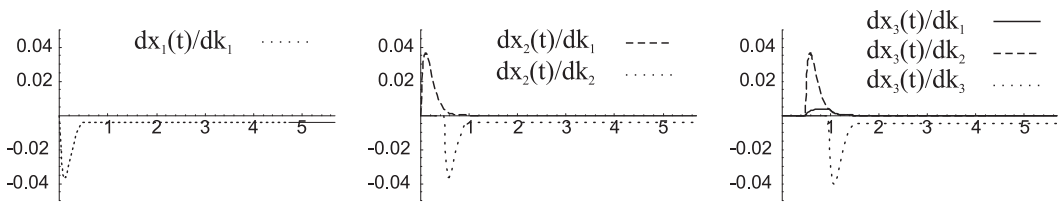


Figure 2. Sensitivity coefficients.

5 Conclusions

Our paper supposes the generalization of the optimal control problem that arises when considering a linear unbranched chemical process with n steps. We provide a quasi-analytical solution to the case of n steps by considering the minimization of the operation time and non-equal catalytic efficiencies of the enzymes. Using our closed-form equation for the optimal solution, the sensitivity analysis is very easy to perform.

References

[1] E. Klipp, R. Heinrich, H.G. Holzhutter, *Prediction of temporal gene expression. Metabolic optimization by re-distribution of enzyme activities*, Eur. J. Biochem. **269 (22)**, (2002) 5406–5413.

[2] M. Bartl, P. Li, S. Schuster, *Modelling the optimal timing in metabolic pathway activation-Use of Pontryagin's Maximum Principle and role of the Golden section*, BioSystems **101** (2010) 67–77.

[3] D. Oyarzun, B. Ingalls, R. Middleton, D. Kalamatianos, *Sequential activation of metabolic pathways: a dynamic optimization approach*, Bull. Math. Biol. **71(8)** (2009) 1851–1872.

 ISBN: 978-84-616-9216-3

[4] L. Bayon, J.M. Grau, M.M. Ruiz, P.M. Suarez, *Optimal control of a linear unbranched chemical process with steps: the quasi-analytical solution*, J. Math. Chem. **52(4)**, (2014) 1036-1049.

[5] T. Turanyi, *Sensitivity analysis of complex kinetic systems. Tools and applications*, J. Math. Chem. **5(3)** (1990) 203-248.

[6] M. Llorens, J.C. Nuno, Y. Rodriguez, E. Melendez-Hevia, F. Montero, *Generalization of the theory of transition times in metabolic pathways: a geometrical approach*, Biophys. J. **77(1)** (1999) 23–36.

[7] R. Vinter, *Optimal Control, Systems & Control: Foundations & Applications*, Birkhäuser Boston, Inc., Boston, MA, 2000.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7 July, 2014.

Improving an autotuning engine for 3D Fast Wavelet Transform on GPU

Gregorio Bernabé[1], Javier Cuenca[1], Luis Pedro García[2] and Domingo Giménez[3]

[1] *Computer Engineering Department, University of Murcia*

[2] *Servicio de Apoyo a la Investigación Tecnológica, Technical University of Cartagena*

[3] *Computer Science and Systems Department, University of Murcia*

emails: gbernabe@ditec.um.es, javiercm@ditec.um.es, luis.garcia@sait.upct.es, domingo@um.es

Abstract

In this paper, we present an enhanced auto optimization method to run the 3D-Fast Wavelet Transform (3D-FWT) on the different NVIDIA GPU devices in a system. The proposed method automatically selects the optimal block size and the number of streams in order to reduce the total execution time, obtaining performances very close to the optimal and decreasing the number of evaluations needed.

Key words: Autotuning engine, 3D-FWT, manycore GPUs, CUDA, streams.

1 Introduction

Over the last decade, general-purpose GPU computing [1][2] has evolved from being something of a curiosity into an extremely popular and immensely powerful HPC platform. There are currently many APIs for programming GPUs, each with their advantages and disadvantages, but getting optimal performance from the GPU is still a challenging task that requires repetitive manual tuning.

NVIDIA has been a driving force in this process through the development of GPU-based hadware for general computation and the parallel development of the CUDA programming model [3].

The emergence of the Fermi GPU and the appearance of the new Kepler GPU [4] in the market have been crucial for the incorporation of streams as a key factor in codes. A stream

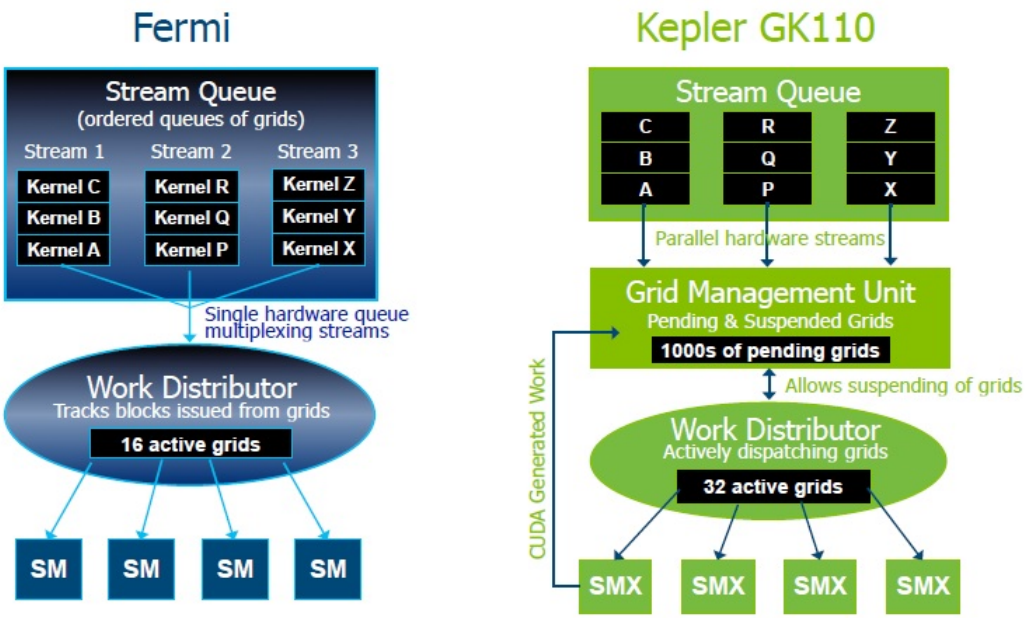


Figure 1: How streams are treated in Fermi and Kepler GPU architectures [4]

is a sequence of operations that execute in issue-order on the GPU. In the programming model used for concurrency, CUDA operations in different streams may run concurrently, or operations from different streams may be interleaved. One of the most difficult challenges for the GPU architecture is finding an optimal scheduler to manage the workload composed of different streams in a GPU.

The Fermi GPU architecture allows a concurrency execution of up to 16 streams, but there is a single hardware queue and the streams must be multiplexed and serialized, as we can observe on the left in Figure 1. This dependence can be alleviated by the rearrangement of the kernels but this task becomes complex and the performance decreases as the complexity of the programs increases.

The new Kepler GPU architecture introduces Hyper-Q, which enables up to 32 hardware queues (Figure 1, right), allowing great flexibility to improve the performance without modifications to the source codes. In Kepler there are no inter-streams dependencies and each stream is managed by its own hardware queue. Streams may proceed from a single CUDA program or from other places in different MPI processes or POSIX threads (pthreads). In this way, the concurrency is natural and does not require preprocessing. As the number of cores increases, Hyper-Q gets more powerful, becoming a key factor in the scalability of future generations of GPUs.

In [5][6] we proposed an autotuning architecture for the 3D-FWT on a cluster of mul-

 ISBN: 978-84-616-2723-3

ticores+GPUs. The method analyzes the different nodes of the cluster, and detects the number and type of CPUs and GPUs, the computer performance of the GPUs and CPUs and the bandwidth of the interconnection network. The autotuning engine computes the proportions at which the different video sequences are divided among the nodes in the cluster.

Here, we present an enhanced optimization engine based on an optimal selection of the number of streams and the block size for the 3D-FWT on each NVIDIA GPU available in the system. A reduced number of evaluations are made in order to select the minimum execution time.

The rest of this paper is organized as follows. Section 2 summarizes the implementation of the 3D-FWT in CUDA. Section 3 analyses the incorporation of the availability of streams to the 3D-FWT CUDA implementation. In section 4, the enhaced optimization technique for a 3D-FWT on a single GPU system is described. Experimental results of this method are discussed in Section 5. Finally, section 6 summarizes and introduces future work.

2 Parallelization on a manycore GPU in CUDA and OpenCL

This section briefly describes the characteristics of the software used in the experiments. Our 3D-FWT implementations in CUDA and OpenCL [7] are based on the CUDA algorithm described in [8]. We use simple source-to-source translation to convert the kernels of the implementation of 3D-FWT on CUDA to OpenCL; although there are some differences between CUDA and OpenCL in terminology, the model is similar and it was easy to transform the kernels. Our 3D-FWT implementation in CUDA and OpenCL consists of three main steps:

1. The *host* (CPU) allocates in the memory the first four video frames coming from a .pgm file.

2. The first four images are transferred from main memory into video memory. The 1D-FWT is then applied to the first four frames over the third dimension to obtain two frames for the detailed and reference videos.

3. The 2D-FWT is applied to the frame belonging to the detailed video, and, subsequently, to the reference video. Results are then transferred back to the main memory.

The whole procedure is repeated for all the input frames, adding two frames in each iteration. Figure 2 summarizes how the entire process is implemented. In each iteration, two frames are copied, to the first or the second half, depending on the iteration number. In particular, the first iteration copies frames number 0, 1, 2 and 3 to obtain the first detailed and reference video frames; the second iteration involves frames 2, 3, 4 and 5 to obtain the

 ISBN: 978-84-616-2723-3

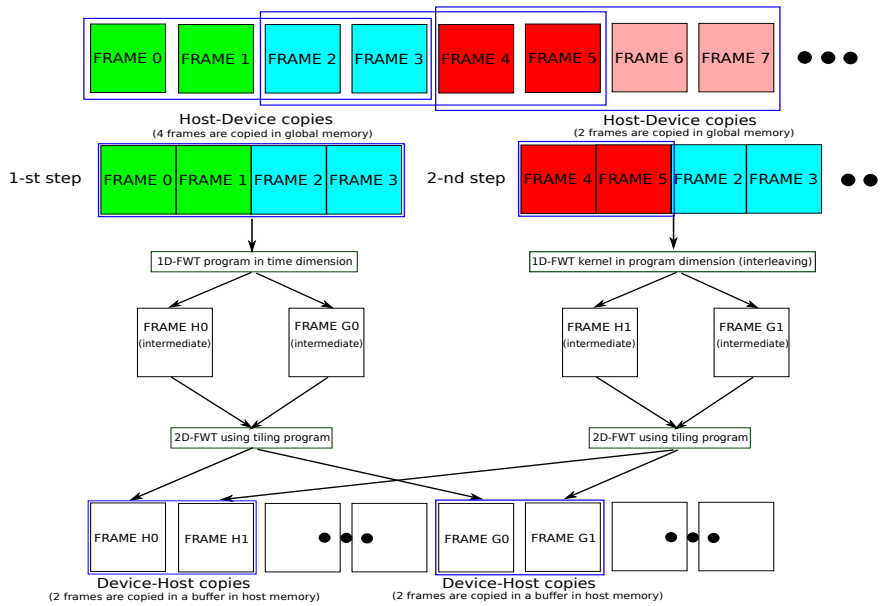


Figure 2: How 3D-FWT is implemented in OpenCL and CUDA using interleaved accesses to video frames.

second detailed and reference video frames, and so on. Note that frames 4 and 5 occupy the memory formerly assigned to frames 0 and 1, which requires an interleaved access to frames in the second iteration.

3 3D-FWT on CUDA with streams

Up to now the main key factor in any CUDA implementation in order to reduce the execution time has been the selection of the optimal block size. The possibility of the incorporation of streams to the codes is another option to improve the execution times, but the selection of the best number of streams is not a trivial task because it can depend on both the hardware (the specific GPU used) and the software (the routine to execute). We have included in our 3D-FWT implementation in CUDA, presented in section 2, the availability for using a different number of streams. Figure 3 shows execution times for the 3D-FWT to process 256 frames of 512×512 pixels on an NVIDIA Fermi Tesla C2075 (to the left) and an NVIDIA Kepler Tesla K20c (to the right). The figure presents different series grouped by typical block sizes and number of streams. Based on a previous work [9], we consider Daubechie's W_4 mother wavelet [10] as an appropriate baseline function. This selection determines the access pattern to memory for the entire 3D-FWT process and requires four elements to calculate the output. Therefore, this sequence of 256 frames can use a maximum of 64

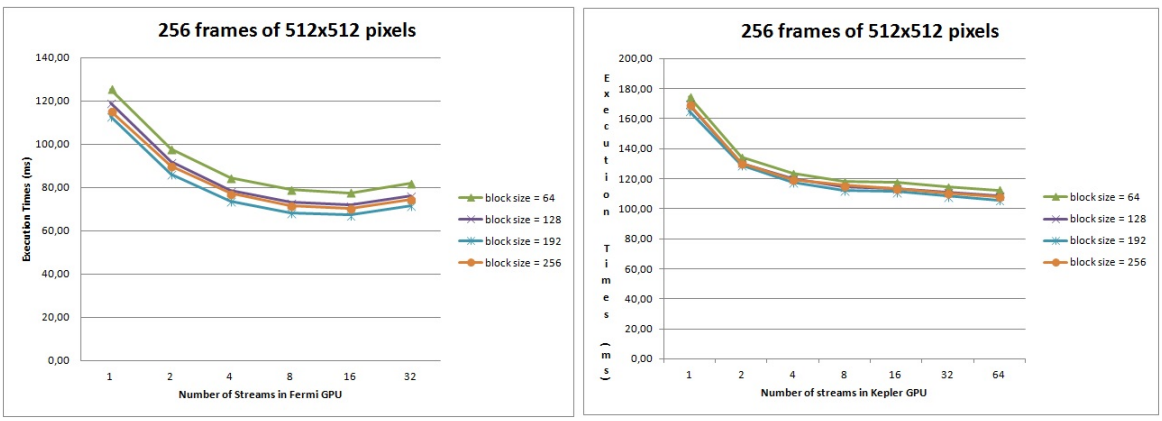


Figure 3: Execution times for 256 frames of 512×512 pixels on Fermi GPU (left) and Kepler GPU (right)

streams, where each stream processes at least 4 frames. A slight difference among execution times is observed for different block sizes in both GPUs. In fact, the maximum difference is about 16% in Fermi GPU and 6% in Kepler GPU. However, the speedups when using several streams with respect to the execution with a single stream is in the range of 1.31 to 1.68 for the Fermi GPU and 1.28 to 1.56 for the Kepler GPU.

In the same way, Figure 4 shows execution times for 3D-FWT to process 256 frames of 1024×1024 pixels on an NVIDIA Fermi Tesla C2075 (to the left) and an NVIDIA Kepler Tesla K20c (to the right). There is small difference in the execution times for different block sizes. The maximum difference is again about 16% in Fermi GPU and 6% in Kepler GPU. The speedups when the number of streams varies is now between 1.28 and 1.58 for the Fermi GPU, and 1.13 and 1.37 for the Kepler GPU.

These results demonstrate a considerable improvement of 3D-FWT execution times with several streams, where an optimal selection of the number of streams is a key factor.

4 An enhanced autotuning engine for the 3D-FWT

We proposed an autotuning engine for the 3D-FWT in [5] and it was adapted to obtain the maximum performance in a hybrid system with several manycore GPUs and multicore CPU components [6]. The number and type of GPUs, and the number of cores in each node is obtained, and the optimization engine computes the workload for each computational component based on the computer performance of the 3D-FWT kernel. In this way, the method automatically decides the quantity of work to scatter among the different platforms

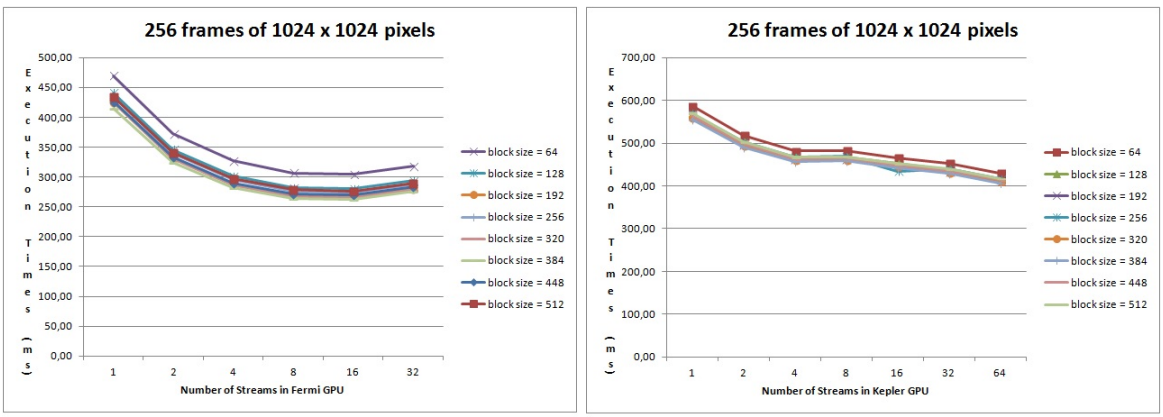


Figure 4: Execution times for 256 frames of 1024×1024 pixels on Fermi GPU (left) and Kepler GPU (right)

of the CPU-GPU system.

This previous proposal performs an optimal selection of the block size in NVIDIA GPUs, but as shown in the previous section, an optimal choice of the number of streams to generate on each GPU device is also an important key. Therefore, an extension of the autotuning methodology is necessary. The improved autotuning method for manycore GPU and multicore CPU systems is shown in Algorithm 1. The method performs a fast analysis for NVIDIA GPUs to obtain the best configuration for the GPU and determines the block size and the number of streams through function `f(block, stream)`.

The complete optimization algorithm for `f(block, stream)` is shown in Algorithm 2. In general, `f(block, stream)` performs a search among a set of block sizes jointly with a set of number of streams in order to find the configuration with the minimum execution time. This search process is performed in a similar way as in [11] in order to reduce the number of possible evaluations and, so, the installation time.

In [12], we empirically determined the optimal configuration for the 3D-FWT using the CUDA occupancy calculator and following a simple set of heuristics [13] [14]. Such a configuration consists of a different thread to compute every pair of G and H values in the 1D-FWT. Each thread requires 13 registers, and a block size of n needs $8n+8$ bytes of shared memory. Thus, the number of active thread blocks per multiprocessor requires a number of registers and an amount of shared memory which must not exceed the maximum allowed values for the NVIDIA GPUs. The proposed automatic function is based on the CUDA occupancy calculator, and the routine computes the occupancy of each multiprocessor for the block sizes $(64, 128, 192, 256, 320, 384, 448, 512)$ recommended by the heuristics [13] [14] in order to select all the block sizes that reach at least 60% occupancy of each multiprocessor

Algorithm 1 Automatic optimization of 3D-FWT for manycore GPUs and multicore CPUs systems

```
Detect automatically the available GPUs and CPUs in the system.
for each platform (GPU or CPU) do
    if GPU is based on NVIDIA then
        Select the CUDA implementation of 3D-FWT.
        f(block, stream) calculates automatically the block size and the number of streams.
    end if
    if GPU is based on ATI then
        Select the OpenCL version of 3D-FWT.
        The work-group size is equal to CL_DEVICE_MAX_WORK_GROUP_SIZE.
    end if
    if CPU then
        Select the implementation with tiling and pthreads.
        Fast analysis to obtain the optimal number of threads.
    end if
    Send one sequence to this platform to obtain the computer performance of the 3D-FWT kernel.
end for
Send sequences in a proportion equal to the 3D-FWT kernel computer performance in each GPU and CPU.
```

Algorithm 2 `f(block, stream)`

Require: Sequence of X frames with a resolution of $n \times n$ pixels

```
{Installation_block_sizes_Set} = {Select automatically block sizes with Occupancy of each Multiprocessor >=60%}
{Installation_streams_Set} = {1 ≤ 2^k ≤ X/4}
threshold = 10%
streams_Set = Installation_streams_Set
best_time = MAX
for each block size in Installation_block_sizes_Set do
    for each number_of_streams in streams_Set do
        Time = Execution 3D-FWT(block size, stream)
        if Time ≤ best_time then
            best_time = Time
            best_block_size = block_size
            best_number_of_streams = number_of_streams
        end if
    end for
    Delete number_of_streams from streams_Set if Time > best_time+threshold
end for
return best_time, best_block_size, best_number_of_streams
```

 ISBN: 978-84-616-2723-3

(line 1 of Algorithm 2). Obviously, the routine contains a table with the physical limits for the GPUs, and the limit of the active thread blocks per multiprocessor is the minimum of the maximum warps, registers and shared memory per multiprocessor.

Next, the $Installation\_streams\_Set$ is composed of several number of streams, each a power of two, between 1 and the frames' number of the input sequence divided by four ($X/4$) (line 2). In a first iteration (lines $6-16$), Algorithm 2 selects a first block size and obtains execution times to process a sequence of frames with the 3D-FWT for the number of streams contained in the $streams\_Set$, which is initialized with the $Installation\_streams\_Set$ (line 4). Next, the function evaluates all execution times and obtains the minimum execution time (best_time). If the execution time for a number of streams is greater than the best_time plus a threshold (established to 10% in line 3), the number of streams of this execution is not considered for the next evaluation of block size (line 15). For the next block sizes, the analysis is only done for the number of streams selected in the previous iteration, so reducing considerably the number of evaluations. Finally, the output is the minimum execution time achieved by a block size and a streams' number (line 17).

5 Experiments

In this section, we test the `f(block, stream)` for an NVIDIA Tesla K20 GPU with 2496 cores, an NVIDIA Fermi Tesla C2050 GPU with 448 cores and an old NVIDIA Tesla C870 GPU with 128 cores. We explain in detail the installation phase for the first GPU and the results obtained for the other two GPUs. Results of the execution phase are analyzed for the three GPUs, including a comparison with to a non-expert user and an expert user.

5.1 Installation phase

For an NVIDIA Tesla K20 GPU and a sequence of 256 frames of 1024×1024 pixels the `f(block, stream)` is executed. In the installation stage $Installation\_block\_sizes\_Set = \{128, 192, 256, 320, 384, 448, 512\}$, $Installation\_streams\_Set = \{1, 2, 4, 8, 16, 32, 64\}$, and $threshold = 10\%$. Table 1 shows execution time of 3D-FWT for different block sizes and number of streams. Moreover, the last column of this Table shows the best_time plus 10% in each iteration with a different block size. For the first block size (128), $streams\_Set$ is equal to $\{1, 2, 4, 8, 16, 32, 64\}$. At the beginning of the second iteration, the best_time is 414.73 msecs., therefore $streams\_Set$ is reduced to $\{16, 32, 64\}$ for the next block size (192). In the following four iterations the $stream\_Set$ is maintained at $\{16, 32, 64\}$. In the last iteration, as the best_time is 405.37 msecs., which has been obtained by a block size of 384 and 64 streams, the $stream\_Set$ is reduced to $\{32, 64\}$. In this situation, the optimization engine obtains the minimum execution time, reducing the number of executed evaluations from the 56 total possible (7 tests for each block size in the the Installation_block_sizes_Set block sizes plus 7 evaluations of the block size 64 previously discarded in this phase) to 24.

 ISBN: 978-84-616-2723-3

Our enhanced automatical method achieves the optimal configuration with a block size of 384 and 64 streams in 10.61 secs.

Table 1: Execution times (msecs.) of `f(block, stream)` for an NVIDIA Tesla K20 GPU

block size/streams	1	2	4	8	16	32	**64**	best_time+10%
128	568.69	501.68	466.32	467.52	451.67	438.49	414.73	456.20
192					447.05	433.93	410.18	451.20
256					434.33	439.08	415.91	457.50
320					444.47	431.28	409.68	450.65
384					442.41	428.84	**405.37**	445.91
448					448.94	435.84	414.54	445.91
512						438.99	416.87	445.91

The `f(block, stream)` is executed for an NVIDIA Fermi Tesla C2050 GPU with 448 cores and a sequence of 128 frames with a resolution of 2048×2048 pixels. Table 2 shows the execution time of 3D-FWT for different block sizes and number of streams. The last column shows the best_time plus 10% in each iteration with a different block size. The best_time is 551.35 msecs. achieved by 384 block size and 8 streams. In this example, the optimization engine obtains the minimum execution time, executing 50.00% of the total evaluations in 14.33 secs.

Table 2: Execution times (msecs.) of `f(block, stream)` for an NVIDIA Fermi Tesla C2050 GPU

block size/streams	1	2	4	**8**	16	32	Best_Time+10%
128	863.61	681.24	602.13	576.10	592.44	659.70	633.71
192			583.30	557.59	574.58		613.35
256			589.62	564.49	581.43		620.94
320			590.83	566.25	583.10		622.88
384			576.60	**551.35**	567.44		606.49
448			581.87	557.56	574.57		606.49
512			594.44	569.55	586.08		606.49

The `f(block, stream)` is executed for an NVIDIA Tesla C870 GPU and a sequence of 64 frames of 1024×1024 pixels. Table 3 shows the execution time of 3D-FWT for different block sizes and numbers of streams. For each block size, the best_time plus 10% is also shown. The best_time, 201.81 msecs., which matches with the optimal configuration, is obtained for the 192 block size and 16 streams. The execution installation time is 5.32 secs. for this GPU.

5.2 Execution phase

In the installation phase, for a video sequence of 10 hours with 25 frames per second and a resolution of 1024×1024 pixels, split into groups of 256 frames, the proposed autotuning

Table 3: Execution times (msecs.) of `f(block, stream)` for an NVIDIA Tesla C870 GPU

block size/streams	1	2	4	8	**16**	Best_Time+10%
64	220.89	220.27	219.71	218.58	216.46	238.11
128	209.59	209.15	208.44	207.33	205.17	225.69
192	206.08	205.56	204.93	203.86	**201.81**	221.99
256	212.41	212.11	211.24	210.16	207.96	221.99
512	224.30	223.90	222.92	221.66	218.91	221.99

engine selects block size 384 and 64 streams for the NVIDIA Tesla K20 GPU. In this way, $900,000$ frames are processed in 23.75 minutes by a non-expert user using the autotuning engine, whereas, without our method, this user, who has no knowledge to properly select the block size and the number of streams, would spend approximatelly 34.38 minutes (selecting 64 as the block size and 1 as the number of streams). On the other hand, an expert user, who selects the optimal block size and establishes the number of streams to 32, which is the number of hardware queues in a Tesla K20 GPU, would take 25.13 minutes. Therefore, speedups of 1.45 and 1.06 are obtained with regard to a non-expert user and an expert user, respectively.

For the NVIDIA Fermi Tesla C2050 GPU, a video sequence of 10 hours with 25 frames per second and a resolution of 2048×2048 pixels, split into groups of 128 frames, is processed in 64.61 minutes with our enhanced automatical method, while a non-expert user would spend 107.73 minutes. An expert user, who selects 1 stream or 16 streams, which are the numbers of theoretical streams allowed in concurrency and the hardware queues in a Fermi Tesla C2050 GPU, would take 97.48 or 66.50 minutes. Speedups of our optimization engine are 1.67 with regard to a non-expert user and 1.51 or 1.02, depending on the selection of 1 or 16 steams by an expert user.

For the NVIDIA Tesla C870 GPU, the $900,000$ frames, split into groups of 64 frames, are processed by our autotuning engine in 47.30 minutes, while a non-expert user would spend 52.57 minutes and an expert user 48.30 minutes. Speedups of our proposal are 1.11 and 1.02, respectively.

6 Conclusions and future work

We propose an extension of a previously proposed optimization engine to run the 3D-FWT kernel automatically on integrated systems with different platforms such as multicore CPU and manycore GPUs. This extension is based on an optimal selection of the block size and the number of streams for an implementation of the 3D-FWT in CUDA. The autotuning method perfoms a fast analysis for NVIDIA GPUs to obtain the best configuration which achieves the minimum execution time for the GPU and determines the block size and the number of streams, reducing the number of possible evaluations.

Our proposed method obtains speedups of up to 1.45 for the NVIDIA Tesla K20, 1.67 for the Fermi Tesla C2050 and 1.11 for the Tesla C870 with regard to a user with no knowledge in selecting the optimal block size and the number of streams. For expert users, who select the optimal block size and know the architecture of the GPUs, the autotuning engine achieves speedups ranging from 1.02 to 1.51 for the three GPUs.

We are to integrating the extension proposed here in the autotuning architecture for the 3D-FWT on a cluster of multicores+GPUs previously proposed [5][6]. A comparison between the new method and the previous proposal will be made. The methodology described in this paper is applicable to other complex compute applications. Following this, non expert users can obtain good performances in other applications. Our work is part of the development of an image processing library oriented toward biomedical applications, allowing users the efficient automatic execution of different routines.

Acknowledgements

This work was supported by the Spanish MINECO, as well as by European Commission FEDER funds, under grant TIN2012-38341-C04-03.

References

[1] D. Manocha, General-Purpose Computation Using Graphic Processors, IEEE Computer 38 (8) (2005) 85–88.

[2] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. E. Lefohn, T. J. Purcell, A Survey of General-Purpose Computation on Graphics Hardware, Computer Graphics Forum 26 (1) (2007) 80–113.

[3] CUDA Zone maintained by NVIDIA, http://www.nvidia.com/object/cuda.html (2009).

[4] NVIDIA, Whitepaper NVIDIA's Next Generation CUDA Compute Architecture: Kepler GK110, http://www.nvidia.com/content/pdf/kepler/nvidia-kepler-gk110-architecture-whitepaper.pdf (2012).

[5] G. Bernabé, J. Cuenca, D. Giménez, Optimization techniques for 3D-FWT on systems with manycore GPUs and multicore CPUs, in: International Conference on Computational Science, 2013.

[6] G. Bernabé, J. Cuenca, D. Giménez, Optimizing a 3D-FWT code in heterogeneous cluster of multicore CPUs and manycore GPUs, in: 25th International Symposium on Computer Architecture and High Performance Computing, 2013.

[7] The Khronos Group, The OpenCL core API specification, http://www.khronos.org/registry/cl (2011).

[8] J. Franco, G. Bernabé, J. Fernández, M. Ujaldón, Parallel 3D Fast Wavelet Transform on manycore GPUs and multicore CPUs, in: 10^{th} International Conference on Computational Science, 2010.

[9] G. Bernabé, J. González, J. M. García, J. Duato, A New Lossy 3-D Wavelet Transform for High-Quality Compression of Medical Video, in: Proceedings of IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 2000, pp. 226–231.

[10] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, 1992.

[11] J. Cámara, J. Cuenca, D. Giménez, L. P. García, A. Vidal, Empirical installation of linear algebra shared-memory subroutines for auto-tuning, International Journal of Parallel Programming 42 (2014) 408–434.

[12] J. Franco, G. Bernabé, J. Fernández, M. E. Acacio, A Parallel Implementation of the 2D Wavelet Transform Using CUDA, in: 17^{th} Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, 2009.

[13] NVIDIA Tutorial at PDP'08, CUDA: A New Architecture for Computing on the GPU (February 2008).

[14] NVIDIA Computing Webinars, NVIDIA GPU Computing Webinars Further CUDA Optimization (April 2009).

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Competition among invasive and native species: the case of European and mountain hares

Alex Berruti[1], Valentina La Morgia[2], Ezio Venturino[1] and Simone Zappalà[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino, via Carlo Alberto 10, 10123 Torino, Italy*

[2] *Istituto Superiore per la Protezione e la Ricerca Ambientale, via Ca' Fornacetta 9, 40064 Ozzano Emilia (BO), Italy*

emails: berruti.alex@gmail.com, valentina.lamorgia@isprambiente.it, ezio.venturino@unito.it, simone.zappala.00@gmail.com

Abstract

A model for the interactions of three hare species in the north-west of Italy is proposed, based on ideas borrowed from the concept of herd behavior for modeling their interactions. The possibility of the coexistence of all the species in the system through persistent oscillations is discovered.

Key words: population dynamics, invasion, interspecific competition
MSC 2000: AMS codes 92D25, 92D40

1 Introduction

In this paper we consider the important problem of invasive speciers, which has been studied already in [5], but in a much different setting. Moving from the case of American grey squirrel (*Sciurus carolinensis* Gmelin, 1788) invading Europe and slowly outcompeting the native red squirrel (*S. vulgaris* (Linnaeus, 1758)), we rather consider the problem of the interplay of two species of hares in northern Italy, the the European hare, *Lepus europaeus* (Pallas, 1778) and the indigenous mountain hare, *Lepus timidus* (Linnaeus, 1758).

While among squirrels we have studied the competition for tree seeds, which are the main source of food for the squirrels, we consider here instead mainly the fight for the

territory. In fact the European hare has been settling in the plains of the north, slowly pushing away the native hare. Now the latter thrives at higher ground on the mountains and partly in segregated areas, even from its own similar, with the territory in between occupied by the invasive population. In fact, all these populations are mainly stantial. Furthermore there is the phenomenon of coupling between these populations at the boundary of the territories that they occupy to be taken into account. This gives rise to a hybrid species, which in addition reproduces on its own. Figure 2 contains a map the current situation in part of Piedmont (NW Italy).

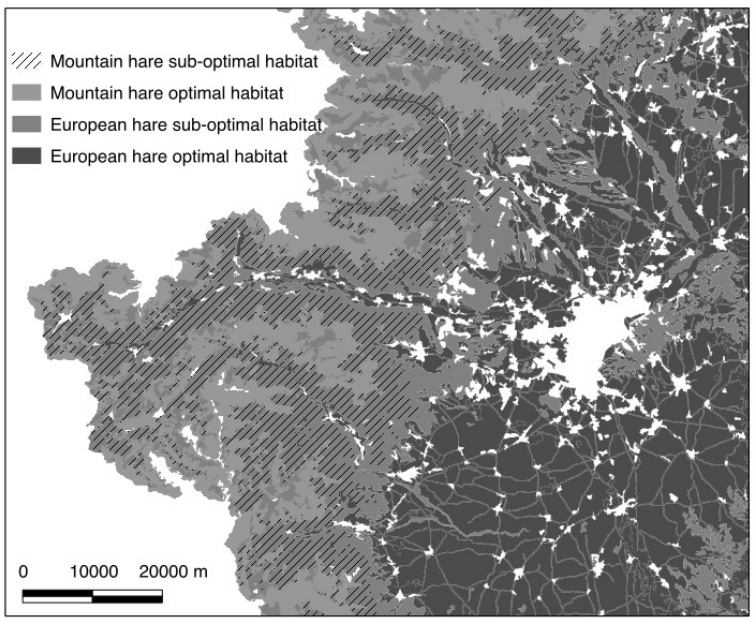


Figure 1: Suitable habitats for the mountain and European hares in the Western Italian Alps. The map is showing part of the Turin province, where the optimal habitat for the European hare is mainly located in lowlands and valley bottoms (dark grey areas). On the contrary, the optimal habitat for the mountain hare is located at the highest altitudes (light grey areas). In-between, the medium grey, dashed areas actually identifies the zones where sub-optimal habitats of the two species overlap (and where hybrids presumably thrive). The habitat suitability is based on data provided by [3].

We propose a model to investigate the relationships among these species, and possibly understand their future evolution. The system is based on the ideas first proposed in [1, 2]. The approach indeed relies on the fact that the hares occupy different territories and therefore among them the interactions can only occur at the border of their respective habitats. This mathematically is modeled via square roots of the populations. In fact, if

 ISBN: 978-84-616-9216-3

we assume that they are distributed over the territory, so that the number of individuals occupying the outermost positions is proportional to the square root of the total population size. This approach differs quite sensibly from the older concept of group defence expounded in [6], which uses suitable assumptions on the shapes of the response functions, i.e. the interactions terms. Note that the square root idea has been exploited also earlier in the context of plankton dynamics, [4].

In the present context, we use these root terms only for accounting for the interspecific interactions among the three hare populations.

The presentation is organised as follows. In the next Section we introduce the dynamical system and its simplified version. Section 3 contains the equilibria analysis. A brief discussion of the results concludes the investigation.

2 The model

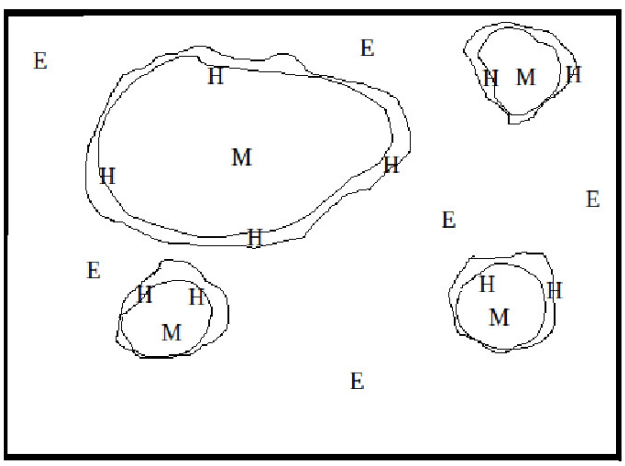


Figure 2: A schematic illustration of the territory in which the three populations thrive, whose interactions we intend to model. The European E hare occupies most of the flatland, leaving the higher ground to the indigenous mountain hare population M. At the elongated borderlines we find the hybrid population H.

We consider here the three hare populations, the European hare E, also known as brown hare o European brown hare, *Lepus europaeus* (Pallas, 1778), the mountain hare M, *Lepus timidus* (Linnaeus, 1758), and the hybrid hare H.

We summarize the basic assumptions underlying the ecosystem in consideration. For the convenience of the reader, Figure 2 contains a schematic picture of the situation that we want to model. The hare is essentially a "stantial" population. The E and M populations

occupy different neighboring habitats so that they hardly directly compete. On the margins of the territory occupied by E there are patches, usually at higher altitude, where the M thrive. At the common, much elongated, boundaries of these different habitats for E and M the two populations come in contact and they originate new individuals, of hybrid type, H. The latter can further reproduce among themselves, as well as by coupling with both confining populations E and M.

The fact that intermingling of these populations occur only on the boundary of the regions of interest suggests us to model their interactions via square root fuctions, as recently proposed in models for prey herd behavior, [2]. Recalling here the assumptions that led to this herd defense model formulation, as they underlie also the ecosystem we are considering here, we model the interactions occurring on the boundaries of their respective territories as follows:

$$\begin{aligned} \frac{dE}{dt} &= rE\left(1-\frac{E}{K}\right) - \widetilde{a}\sqrt{E}\sqrt{M} - b\sqrt{E}H, \\ \frac{dM}{dt} &= sM\left(1-\frac{M}{L}\right) - \widetilde{c}\sqrt{E}\sqrt{M} - e\sqrt{M}H, \\ \frac{dH}{dt} &= qH - nH^2 + \left[q_M\sqrt{M} + q_E\sqrt{E} - \left(g\sqrt{E} + f\sqrt{M}\right)\right]H + \widetilde{w}\sqrt{EM}. \end{aligned} \tag{1}$$

The first equation describes the evolution of the European hare. It grows logistically with net reproduction rate r and carrying capacity K, and competes for resources with the mountain hare population at rate $\widetilde{a}$. Since the environments in which these populations live are only in part overlapping, the interactions are considered as if they were occurring only for the animals living on the border of each environment. This is modeled via the use of the square roots of these populations. The population E further competes with the hybrid population at rate b. Note that again not the whole population E is involved in the interspecific competition, but only the fraction of the population that resides on the boundary, which is expressed once more by the square root term, [2]. Note also that the root does not involve the hybrid population, because as we said, this population lives at the intersection of the territories where M and E live, and this boundary zone can be thought of as a thin and possibly long stripe. Therefore it resembles a one dimensional manifold, so that it essentially coincides with its boundary. Therefore H is not really distributed over a two-dimensional domain, but rather on a set of essentially long one-dimensional patches.

The second equation models the dynamics of the mountain hare. Again we have logistic behavior, with net reproduction rate s. The carrying capacity L incorporates all the possible fragmented habitats, with no communications in between, where the mountain hares find refuge, after the invasion of the European hares. Competition in this case occurs once on the neighboring territories where the European hare thrives, at rate $\widetilde{c}$. Once more, this occurs among the individuals that occupy the boundary of the environment of M that borders

 ISBN: 978-84-616-9216-3

with the patches in which the E's thrive, and the interaction involves only the part of the M and E populations that are close to these boundaries. This is again expressed by the square root terms. The interaction with the hybrid hare occurs again on the border, giving rise once more to the square root term in M, at rate e.

For the hybrid hare instead, third equation, the dynamics is a bit more involved. Reproduction occurs again in a kind of logistic fashion, with net birth rate q and intraspecific competition rate n. But hybrid hare offsprings are generated also via coupling of the hybrid individuals with both mountain and European hares, at rates q_M and q_E respectively, keeping in mind that again this can occur only on the common boundaries of their respective habitats. In addition, although we can assume it to be a seldom occurrence, direct couplings of European and mountain hare produce hybrid newborns, at rate $\widetilde{w}$. Since these events take place in the few locations where possibly these two populations interact, we assume that the individuals involved are those living at the outskirt of their respective habitats, and therefore model the couplings via the square root terms of both M and E populations. Competition of the hybrid hare occurs with both E and M individuals, since the environment in which the hybrid hare lives borders both territories where E and M thrive. The latter populations are therefore only "marginally" interested, as mentioned above, i.e. only their fractions on the border of their habitats are involved, once again justifying the square root terms.

Now $\widetilde{a}$, $\widetilde{e}$ and $\widetilde{w}$ as mentioned are very low rates, which actually get smaller and smaller the higher the population H grows, since the latter represents a barrier among the M and E populations, thereby diminishing the mutual interaction between the two stantial populations that can occur only on the interface of their respective territories. In fact if the M and E populations are separated by the habitat of the H's, their mutual interactions are impossible. Thus these are not really constant coefficients. It is the size of the hybrid population at the interface that allows or prevents the interactions of E's and M's. Therefore the parameters $\widetilde{a}$, $\widetilde{c}$ and $\widetilde{w}$ must be rather functions of H. More specifically, they must be decreasing functions of H. In fact, the smaller the H population is, the greater the possibility of direct encounters between E's and M's is and vice versa. We assume therefore the functional forms of Holling type II terms, as follows

$$\widetilde{a}(H) = \frac{a}{m+H}, \quad \widetilde{c}(H) = \frac{c}{m+H}, \quad \widetilde{w}(H) = \frac{w}{m+H}.$$

Alternatively, one could simply assume that direct interactions among European and mountain hares are impossible, i.e. for simplicity assume that $\widetilde{a} = 0$, $\widetilde{e} = 0$, $\widetilde{w} = 0$. The model (1) would then be rewritten as

$$\frac{dE}{dt} = rE\left(1 - \frac{E}{K}\right) - b\sqrt{E}H \tag{2}$$

 ISBN: 978-84-616-9216-3

$$\begin{aligned}\frac{dM}{dt} &= sM\left(1-\frac{M}{L}\right)-e\sqrt{M}H\\ \frac{dH}{dt} &= qH-nH^2+\left[q_M\sqrt{M}+q_E\sqrt{E}-\left(g\sqrt{E}+f\sqrt{M}\right)\right]H.\end{aligned}$$

Let us substitute $P=\sqrt{E}>0$, $U=\sqrt{M}>0$ into (1), to get the singularity-free systems

$$\begin{aligned}\frac{dP}{dt} &= \frac{1}{2}\left[rP\left(1-\frac{P^2}{K}\right)-\frac{a}{H+m}U-bH\right]\\ \frac{dU}{dt} &= \frac{1}{2}\left[sU\left(1-\frac{U^2}{L}\right)-\frac{c}{H+m}P-eH\right]\\ \frac{dH}{dt} &= H\left[q-nH+P\left(q_E-g\right)+U(q_M-f))\right]+\frac{w}{H+m}PU.\end{aligned} \tag{3}$$

and its simplified version

$$\begin{aligned}\frac{dP}{dt} &= \frac{1}{2}\left[rP\left(1-\frac{P^2}{K}\right)-bH\right]\\ \frac{dU}{dt} &= \frac{1}{2}\left[sU\left(1-\frac{U^2}{L}\right)-eH\right]\\ \frac{dH}{dt} &= H\left[q-nH+P(q_E-g)+U(q_M-f))\right].\end{aligned} \tag{4}$$

The Jacobian for the system (3) is

$$J=\begin{pmatrix}\frac{1}{2}r\left(1-\frac{3}{K}P^2\right) & -\frac{1}{2}\frac{a}{m+H} & -\frac{1}{2}b+\frac{a}{2(m+H)^2}U\\ -\frac{1}{2}\frac{c}{m+H} & \frac{1}{2}s\left(1-\frac{3}{L}U^2\right) & -\frac{1}{2}e+\frac{c}{2(m+H)^2}P\\ \frac{w}{m+H}U+(q_E-g)H & \frac{w}{m+H}P+(q_M-f)H & J_{33}\end{pmatrix}, \tag{5}$$

with $J_{33}=q-2nH+(q_E-g)P+(q_M-f)U-\frac{w}{(m+H)^2}PU$, while for (4) it becomes

$$J=\begin{pmatrix}\frac{1}{2}r\left(1-\frac{3}{K}P^2\right) & 0 & -\frac{1}{2}b\\ 0 & \frac{1}{2}s\left(1-\frac{3}{L}U^2\right) & -\frac{1}{2}e\\ (q_E-g)H & (q_M-f)H & \widetilde{J}_{33}\end{pmatrix}. \tag{6}$$

where now $\widetilde{J}_{33}=q-2nH+(q_E-g)P+(q_M-f)U$.

3 Equilibria

3.1 Particular cases

Both (3) and (4) share the origin E_0 and possibly the coexistence $E^*=(P^*,U^*,H^*)$ equilibria. In addition, the simplified system (4) has also the the points $E_1=(0,\sqrt{L},0)$, $E_2=(\sqrt{K},0,0)$, $E_3=(\sqrt{K},\sqrt{L},0)$.

 ISBN: 978-84-616-9216-3

The origin has the eigenvalue $q > 0$ in both models so it is unstable. Also the equilibria E_1 and E_2 are always unstable, in view of the respective eigenvalues $r > 0$, and $s > 0$. For E_3 we find instead $-r < 0$, $-s < 0$ and $q + \sqrt{K}(q_E - g) + \sqrt{L}(q_M - f)$, from which the stability condition follows:

$$q + \sqrt{K}q_E + \sqrt{L}q_M < \sqrt{K}g + \sqrt{L}f. \tag{7}$$

Therefore, the ecosystem survives, and in the simplified case in which direct competition among the indigeneous and the invader populations are avoided, possibly only the subsystem with no hybrid population thrives. But in this case the simplified system becomes less plausible, since in the absence of H direct interactions between E and M should then become possible and therefore should also be modeled in (4), i.e. giving back (3).

The equilibrium E_3 occurs if the competition rates of the hybrid population with both native and invaders are high enough, see (7). This condition is nonempty, as can be seen in Figure 3, obtained for the parameter values $r = 2.6$, $K = 12.6$, $a = 0$, $m = 1$, $b = 1.3$, $s = 1.5$, $L = 20$, $c = 0$, $e = 0.8$, $q = 0.017$, $n = 0.9$, $q_M = 0.02$, $q_E = 0.2$, $g = 0.3$, $f = 0.06$, $w = 0$. Note also that the fact that there are three real eigenvalues for E_3 makes a Hopf bifurcation at this point impossible, i.e. no persistent oscillations can arise in the neighborhood of this equilibrium.

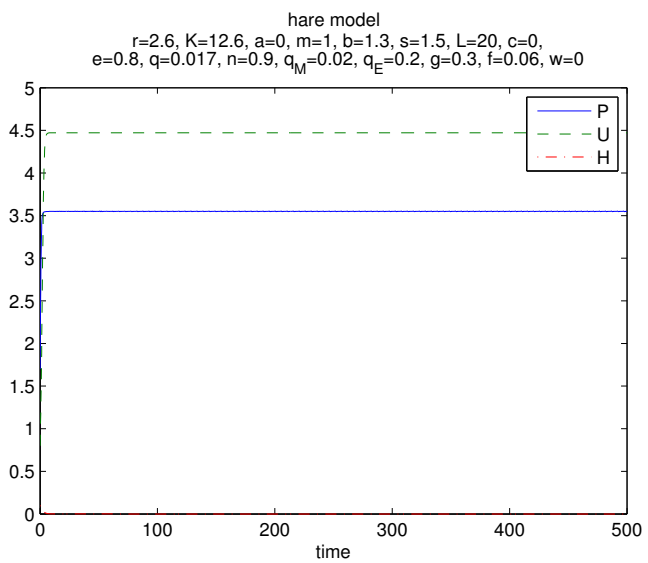


Figure 3: The hybrid hare-free equilibrium E_3 for the model (4) is attained for the following parameter values: $r = 2.6$, $K = 12.6$, $a = 0$, $m = 1$, $b = 1.3$, $s = 1.5$, $L = 20$, $c = 0$, $e = 0.8$, $q = 0.017$, $n = 0.9$, $q_M = 0.02$, $q_E = 0.2$, $g = 0.3$, $f = 0.06$, $w = 0$.

 ISBN: 978-84-616-9216-3

We have also analysed the relationship of the equilibria E_3 and E^* in the simplified model (4). We observed a transcritical bifurcation, which taking as bifurcation parameter q_M, it occurs near the value $q_M^* \approx 0.12$, as it can be observed in Figure 4 for the parameter values $r = 2.6$, $K = 12.6$, $a = 0$, $m = 1$, $b = 1.3$, $s = 1.5$, $L = 20$, $c = 0$, $e = 0.8$, $q = 0.017$, $n = 0.9$, $q_E = 0.2$, $g = 0.3$, $f = 0.06$, $w = 0$. Clearly, the coexistence equilibrium E^* emanates from the hybrid hare-free equilibrium E_3 as its coupling rate with the European hare increases past the threshold value q_M^*.

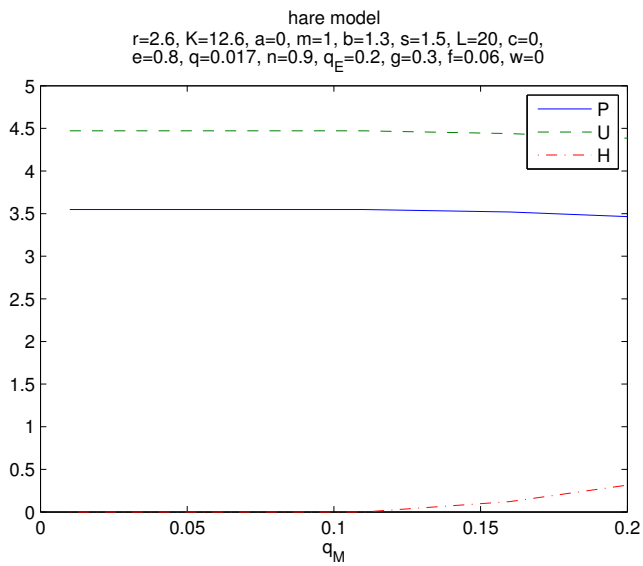


Figure 4: Transcritical bifurcation between E_3 and E^* in model (4). The three hare populations are plotted against the model parameter q_M. It is obtained for the following parameter values: $r = 2.6$, $K = 12.6$, $a = 0$, $m = 1$, $b = 1.3$, $s = 1.5$, $L = 20$, $c = 0$, $e = 0.8$, $q = 0.017$, $n = 0.9$, $q_E = 0.2$, $g = 0.3$, $f = 0.06$, $w = 0$. This picture also clearly shows that the coexistence equilibrium E^* can be attained also for the model (4) for a value of q_M past the threshold $q_M^* \approx 0.12$.

3.2 Coexistence

The analysis of the coexistence equilibrium appears to be much more complicated.

But for the particular case (2), coexistence can be investigated with geometric arguments. Solving the first two equilibrium equations in terms of H, we find that these represent two cylinders in the phase space, with axes parallel to the coordinate axes:

$$H_1 = \frac{r}{b}P\left(1 - \frac{P^2}{K}\right), \quad H_2 = \frac{s}{e}U\left(1 - \frac{U^2}{L}\right), \tag{8}$$

 ISBN: 978-84-616-9216-3

while the third equation gives a plane π,

$$H_3 = \frac{1}{n}\left[q + P(q_E - g) + U(q_M - f)\right]. \tag{9}$$

Clearly the latter has always a portion in the feasible region of the phase plane. The intersection of the two cylinders exists always, and is an arc of a line ℓ joining the origin with the point $(\sqrt{K}, \sqrt{L}, 0)$. Thus, the intersection with the plane π may or may not exist, depending on the inclinations of the latter, i.e. ultimately on its gradient. Moreover, even when the intersection occurs, it could be a single point, or even two points.

A sufficient condition for the existence of a single intersection point can be obtained in the case $q_E < g$, $q_M < f$, by requiring that the abscissae of the intersections of π with the coordinate axes are smaller than the respective rescaled carrying capacities, namely

$$\frac{q}{n(f - q_M)} < \sqrt{L}, \quad \frac{q}{n(g - q_E)} < \sqrt{K}. \tag{10}$$

Whenever conditions (10) hold, the uniqueness of the coexistence equilibrium is guaranteed.

The stability is difficult to assess analytically. But numerical simulations reveal that the coexistence equilibrium can be attained at a stable level, see Figure 5. It is attained for the parameter choice: $r = 2.6$, $K = 12.6$, $a = 0.03$, $m = 1$, $b = 1.3$, $s = 1.5$, $L = 20$, $c = 0.04$, $e = 0.8$, $q = 1.7$, $n = 0.9$, $q_M = 0.2$, $q_E = 0.5$, $g = 0.3$, $f = 0.06$, $w = 0.02$. In view of the fact that this is the only possible equilibrium in case of the full model (3), we conjecture that, whenever locally asymptotically stable, it is also globally asymptotically stable.

We have then tried to investigate also the possibility of existence of persistent sustained oscillations, through repeated simulations involving all parameters, tracing all the population levels for extended parameter ranges and over long periods of time. We were able to find these limit cycles in several situations, presented here in Figures 6-8.

In Figure 9 we report instead a plot in the parameter space to investigate the ranges for which persistent oscillations are possible. We draw in light color the situations for which limit cycles exist, in terms of the reduced parameters $\widetilde{f} = q_M - f$ and $\widetilde{g} = q_E - g$. Note that the light stripes shoot off from the origin, showing that only for these parameter combinations with opposite signs the oscillations can persistently arise.

4 Conclusions

The model presented indicates that it the three species cannot be wiped out, which from the ecological and conservationist viewpoint is a good result. From the invading species viewpoint instead it indicates that the elimination of the European hare has now become impossible by natural means. The possible viable equilibria are the hybrid-hare free point and coexistence. The former is stable if, as remarked, the competition rates of the hybrid

hare with the remaining ones are sufficiently high. But as remarked in the text, the model on which it relies becomes in this situation inadequate. We should then consider this equilibrium as hardly possible. This implies that the hybrid hare is also persistent in this environment.

The three species coexistence is possible not only at a stable level, but also through

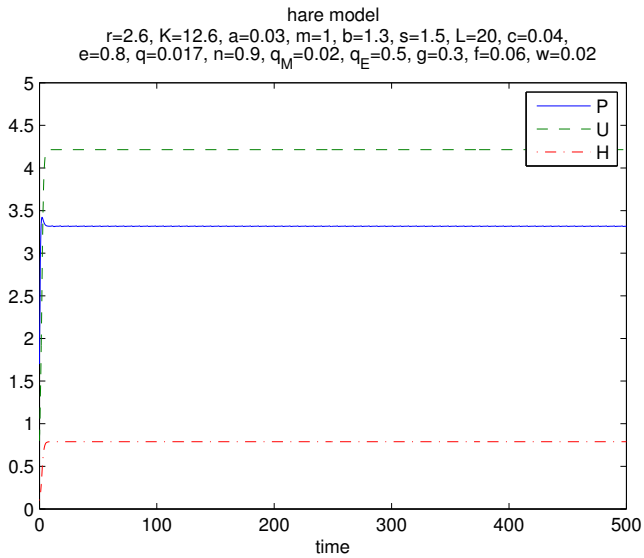


Figure 5: Coexistence equilibrium E^* for the model (3) is attained for the following parameter values: $r = 2.6$, $K = 12.6$, $a = 0.03$, $m = 1$, $b = 1.3$, $s = 1.5$, $L = 20$, $c = 0.04$, $e = 0.8$, $q = 1.7$, $n = 0.9$, $q_M = 0.2$, $q_E = 0.5$, $g = 0.3$, $f = 0.06$, $w = 0.02$.

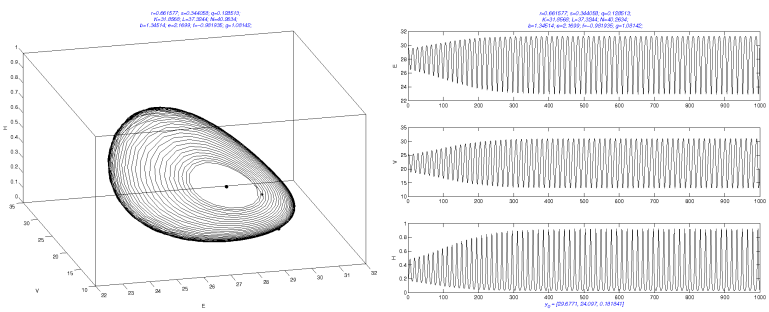


Figure 6: Persistent oscillations for the model (1) are attained for the following parameter values: $r = 0.661577$, $s = 0.344058$, $q = 0.128513$, $K = 31.8568$, $L = 37.3244$, $n = 0.0032$, $b = 1.34514$, $e = 2.1699$, $f = 0.981935$, $g = 1.08142$, $\widetilde{a} = 0$, $\widetilde{c} = 0$, $\widetilde{w} = 0$.

 ISBN: 978-84-616-9216-3

sustained oscillations for all the populations in the ecosystem, as demonstrated by our extended numerical simulations.

References

[1] V. Ajraldi, E. Venturino, *Mimicking spatial effects in predator-prey models with group defense*, Proceedings of the 2009 International Conference on Computational and Mathematical Methods in Science and Engineering, J. Vigo Aguiar, P. Alonso, S. Oharu, E. Venturino, B. Wade (Editors), Gijón, Asturias, Spain, June 30th - July 3rd, 2009, p. 57-66. ISBN 978-84-612-9727-6.

[2] V. Ajraldi, M. Pittavino, E. Venturino, *Modelling herd behavior in population*

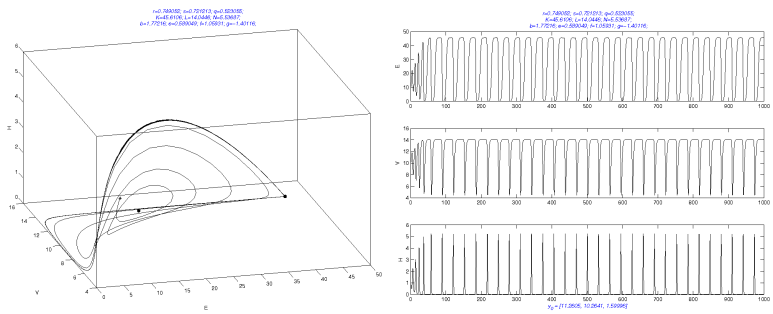


Figure 7: Another case of limit cycles forthe model (1) obtained for the parameter values $r = 0.749052$, $s = 0.721213$, $q = 0.523055$, $K = 45.6106$, $L = 14.0446$, $n = 0.0945$, $b = 1.77216$, $e = 0.589049$, $f = 1.05931$, $g = 1.40116$, $\widetilde{a} = 0$, $\widetilde{c} = 0$, $\widetilde{w} = 0$.

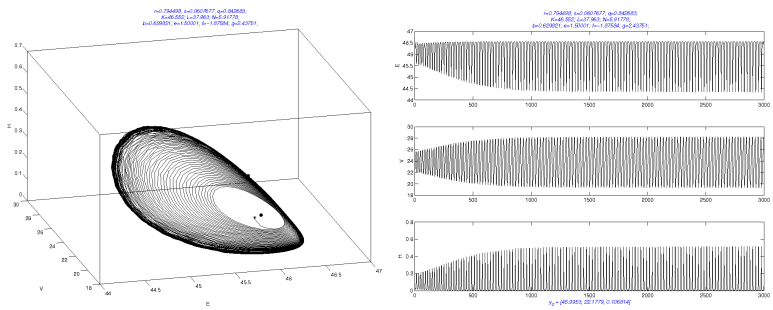


Figure 8: limit cycles for the model (1) for the following parameter values: $r = 0.794498$, $s = 0.0607677$, $q = 0.842683$, $K = 46.552$, $L = 37.963$, $n = 0.1424$, $b = 0.629821$, $e = 1.50001$, $f = 1.87584$, $g = 2.43751$, $\widetilde{a} = 0$, $\widetilde{c} = 0$, $\widetilde{w} = 0$.

systems, Nonlinear Analysis Real World Applications, 12 (2011) 2319-2338.

[3] Boitani L., Corsi F., Falcucci A., Maiorano L., Marzetti I., Masi M., Montemaggiori A., Ottaviani D., Reggiani G., Rondinini C., *Rete Ecologica Nazionale. Un approccio alla conservazione dei vertebrati italiani (National Ecological Network. An approach to Italian vertebrates conservation)*, Università di Roma "La Sapienza", Dipartimento di Biologia Animale e dell'Uomo; Ministero dellAmbiente, Direzione per la Conservazione della Natura; Istituto di Ecologia Applicata (2002).

[4] J. Chattopadhyay, S. Chatterjee, E. Venturino, *Patchy agglomeration as a transition from monospecies to recurrent plankton blooms*, Journal of Theoretical Biology, 253 (2008) 289-295.

[5] A. Gosso, V. La Morgia, P. Marchisio, O. Telve, E. Venturino, *Does a larger carrying capacity for an exotic species allow environment invasion? — Some considerations on the competition of red and grey squirrels*, J. of Biol. Systems 20, No. 3, 221-234, 2012.

[6] H. I. Freedman, G. Wolkowitz, *Predator-prey systems with group defence: the paradox of enrichment revisited*, Bull. Math. Biol., 48 (1986) 493-508.

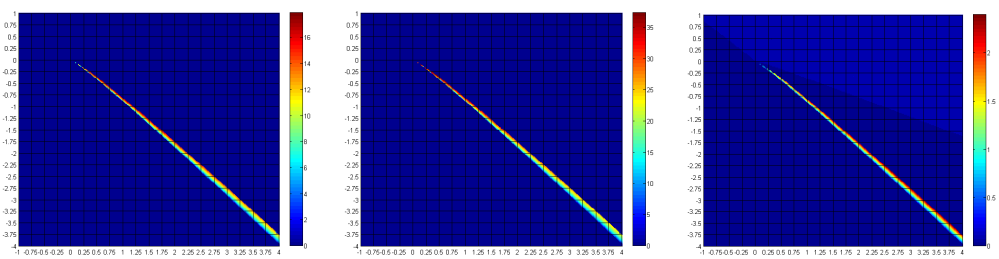


Figure 9: In the parameter space $\widetilde{g} - \widetilde{f}$, where $\widetilde{f} = q_M - f$ and $\widetilde{g} = q_E - g$, the figure shows where the persistent oscillations can be found. It is clearly seen that we need opposite sign values for these parameters in order for these limit cycles to arise, as the light stripes indicating the feasible range shoot off from the origin. Left the population E; Center the population M; Right the population H.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

On how far mosquitos matter in describing dengue fever epidemiology

Juliana Bezerra[1,2], Filipe Rocha[1], Luis Mateus[1], Nico Stollenwerk[1], Paulo Pimenta[2], Eduardo Pessanha[2], Nagila Secundino[2], Jorge Arias[3], Doug Norris[4] and Maira Aguiar[1]

[1] *Centro de Matemática e Aplicações Fundamentais, Universidade de Lisboa, Portugal*

[2] *Laboratory of Medical Entomology, Instituto René Rachou, Belo Horizonte, Brazil*

[3] *Fairfax County Health Department, Virginia, USA*

[4] *Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland Area, USA*

emails: jmt_bezerra@hotmail.com, frocha@ptmat.fc.ul.pt, luisgam1@yahoo.com, nico@ptmat.fc.ul.pt, pfppimenta@gmail.com, edumpessanha@hotmail.com, nagila@hotmail.com, jjoh12@fairfaxcounty.gov, dnorris3@jhu.edu, maira@ptmat.fc.ul.pt

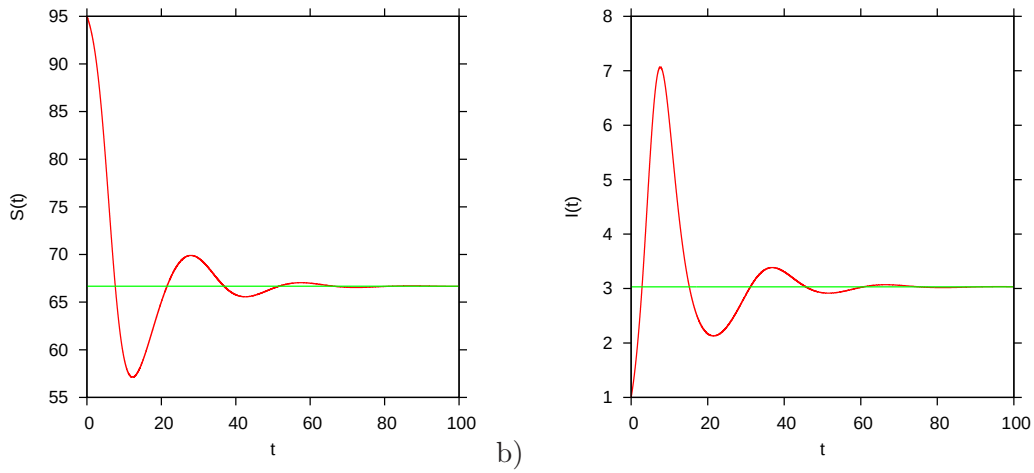
Abstract

We investigate the influence of mosquito dynamics on vector borne diseases, and apply the results to data collected for dengue fever epidemiology in Belo Horizonte. Stochastic enhancement of deterministically transient oscillations is observed.

Key words: Dengue fever, mosquitos, fixed point analysis, dominant fequencies, parameter estimation, stochastic modelling, stochastic enhancement of transient behaviour

1 Introduction

Recently, models with various strains have been investigated to describe the complex behaviour of dengue fever epidemiology [4, 5, 12]. Some advances in understanding the observed fluctuations in dengue fever have been made via simplified models of primary versus secondary infection in simple reinfection models [27]. Here we combine the results from human disease models in dengue fever with mosquito dynamics, as described by simpler models before, an analyse actual data from one city in Brazil concerning mosquito abundance, the ratio of total number of mosquitos versus infected mosquitos and the disease cases of dengue fever in humans.



a) b)

Figure 1: The SIR model described as deterministical ordinary differential equation system. a) The number of susceptibles changing with time, starting from arbitrary initial conditions, oscillates into a stable fixed point, which can be calculated analytically. b) Same for the number of infected.

2 The SIR epidemic process as example

As a first example of an epidemiological process we consider the SIR system with the differential equation system [22]

$$\begin{aligned} \frac{dS}{dt} &= \alpha R - \frac{\beta}{N} S \cdot I \\ \frac{dI}{dt} &= \frac{\beta}{N} S \cdot I - \gamma I \\ \frac{dR}{dt} &= \gamma I - \alpha R \end{aligned} \tag{1}$$

with parameters $\gamma = 1$, $\alpha = 0.1$ and $\beta = 1.5 \cdot \gamma$ and population size $N = 100$. In this parameter region the system shows spiralling into the endemic fixed point $S^* = \frac{\gamma}{\beta} N$, $I^* = \frac{\alpha}{\gamma+\alpha}\left(1 - \frac{\gamma}{\beta}\right) N$, which is typical for many such systems. Hence we investigate this example in more detail, and then transer the results to other models, including models with reinfection and with human and mosquito coupled dynamics. In Fig. 1 we plot the number of susceptibles and of infected against time, and in Fig. 2 we plot the state space, hence $S(t)$ and $I(t)$.

With constant population size $N = S + I + R$ we can reduce the ODE system to a 2 dimensional system and introducing densities $x_1 = S/N$ and $x_2 = I/N$ we obtain the

 ISBN: 978-84-616-9216-3

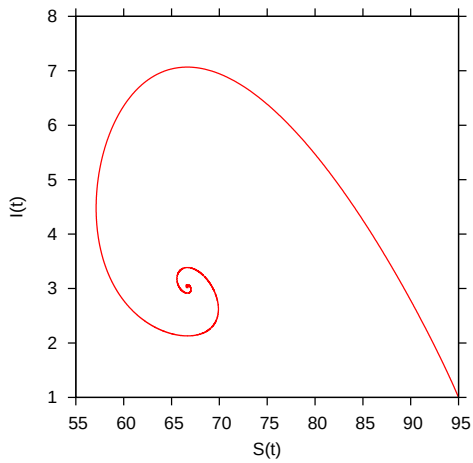


Figure 2: The SIR state space plot for the same system as described in the previous figure shows spiralling into the fixed point.

dynamic equations as

$$\frac{d}{dt}\underline{x} = \underline{f}(\underline{x}) \tag{2}$$

with $\underline{x} = (x_1, x_2)^{tr}$ and $\underline{f}(\underline{x})$ given by $f_1(x_1, x_2) = \alpha \cdot (1 - x_1 - x_2) - \beta x_1 \cdot x_2$ and $f_2(x_1, x_2) = \beta x_1 \cdot x_2 - \gamma \cdot x_2$. We can now linearize around the endemic fixed point $\underline{x}^*$ and calculate for initial conditions $\underline{x}(t_0)$ close to the fixed point the approximate solution from the deviation from the fixed point $\Delta\underline{x} = \underline{x}(t) - \underline{x}^*$. The approximation with linearized dynamics around the endemic fixed point

$$\frac{d}{dt}\Delta\underline{x} = \left.\frac{d\underline{f}}{d\underline{x}}\right|_{\underline{x}^*} \Delta\underline{x} = \begin{pmatrix} -\alpha - \beta x_2^* & -\alpha - \beta x_1^* \\ \beta x_2^* & \beta x_1^* - \gamma \end{pmatrix} \cdot \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \end{pmatrix} \tag{3}$$

and abreviating $\Delta\underline{x} = \underline{y}(t)$ and Jacobian matrix A hence

$$\frac{d}{dt}\underline{y} = A\underline{y} \tag{4}$$

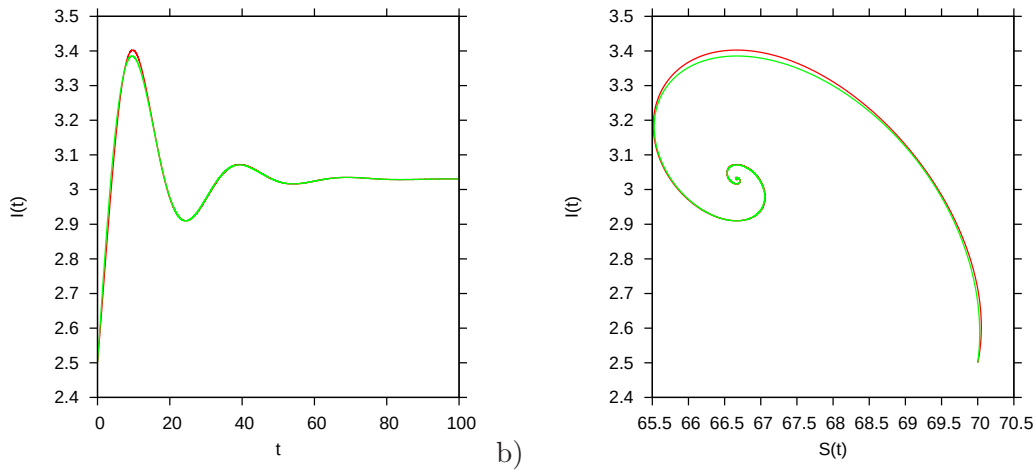
with solution $\underline{y}(t) = e^{A(t-t_0)}\underline{y}(t_0)$ and after eigenvalue/eigenvector decomposition $AT = T\Lambda$, resulting in complex eigenvalues $\lambda_1 = a + i\omega$ and $\lambda_2 = a - i\omega$ with real part a and imaginary part ω as functions of the transition rates,

$$\underline{y}(t) = Te^{\Lambda(t-t_0)}T^{-1}\ \underline{y}(t_0) \tag{5}$$

with time evolution matrix

$$Te^{\Lambda(t-t_0)}T^{-1} \quad = \quad e^{a(t-t_0)} \left(cos(\omega(t-t_0)) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

 ISBN: 978-84-616-9216-3



a) b)

Figure 3: a) Approximation of time series of infected, b) approximation of state space plot by linearized dynamics around the endemic fixed point.

$$+\frac{1}{\omega} sin(\omega(t-t_0)) \cdot \left(\begin{array}{cc} \frac{1}{2}(a_{11}-a_{22}) & a_{12} \\ a_{21} & \frac{1}{2}(a_{22}-a_{11}) \end{array} \right) \Bigg) \tag{6}$$

gives the green line in Fig. 3 as compared to the direct simulation of the SIR system as red line. We will now investigate the same epidemiological system as a stochastic process in order to analyse the influence of noise on the qualitative behaviour of the system.

3 The stochastic system

The SIR epidemiological system can be described as a stochastic process with a time dependent Markov process, also called master equation, as

$$\begin{aligned} \frac{d}{dt}\, p(S,I,t) \;\; = \;\; & \frac{\beta}{N}(S+1)(I-1)\, p(S+1,I-1,t) \\ & +\gamma(I+1)\, p(S,I+1,t) \\ & +\alpha(N-(S-1)-I)\, p(S-1,I,t) \\ & -\left(\frac{\beta}{N}SI+\gamma I+\alpha(N-S-I)\right)\, p(S,I,t) \end{aligned} \tag{7}$$

and can be simulated on a computer via the Gillespie algorithm [13, 14] via exponential waiting times in epidemiological states and then stochastic transitions into other epidemiological states.

 ISBN: 978-84-616-9216-3

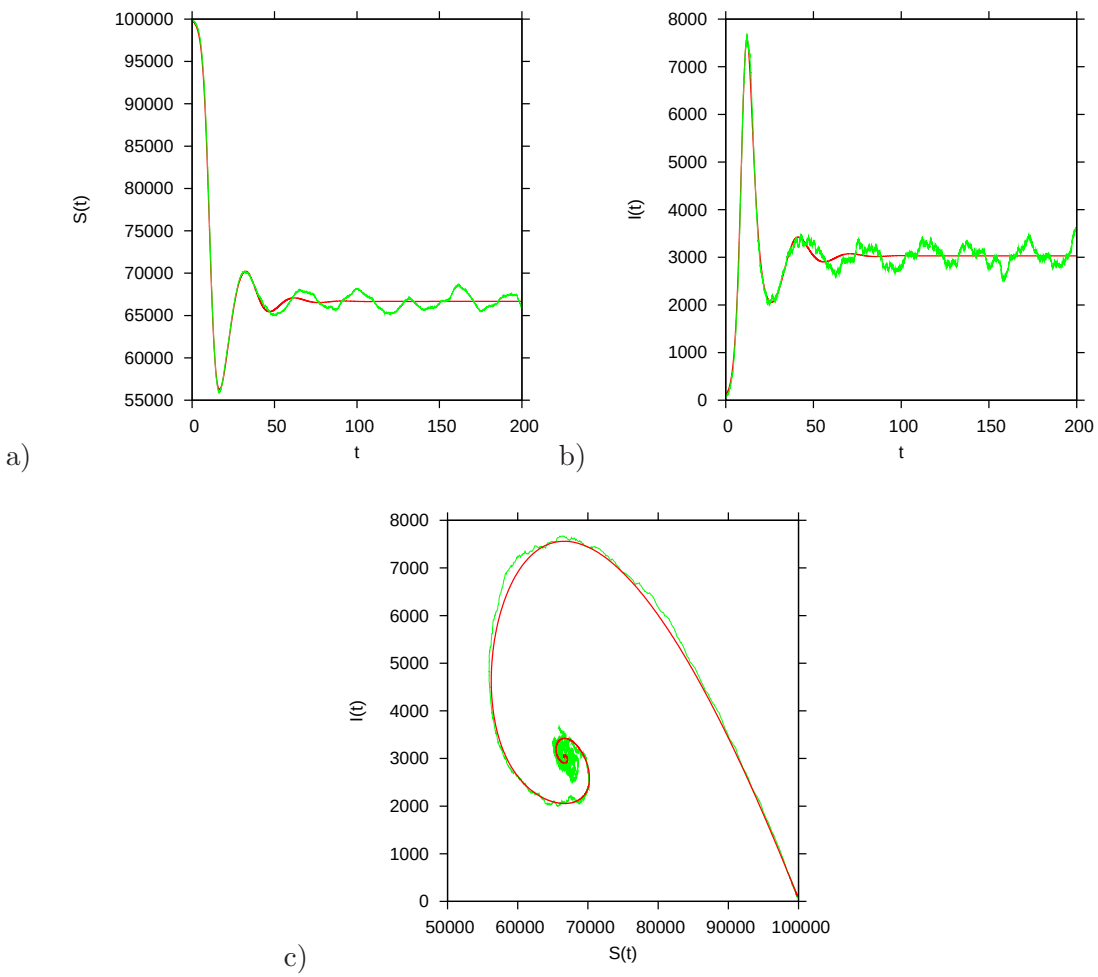


Figure 4: The stochastic SIR system shows continued oscillations around the deterministic fixed point. a) Time series of the susceptibles, b) of infected and c) in state space of the stochastic system in comparison with the deterministic SIR system.

 ISBN: 978-84-616-9216-3

The stochastic system trajectory originally follows closely the deterministic trajectory, Fig. 4, but while the deterministic trajectory spirals into a stable fixed point, the stochastic trajectory still displays oscillation of a similar period length as the transients of the deterministic system. Hence the stochasticity enhances the fluctuations originating from the transients of the deterministic system, a behaviour which can be observed in many population biological systems [24, 25]. It is in epidemiology not only typical for the SIR system but also holds in more extended system, like the reinfection model which we will show now and also in a more extended model for vector borne diseases, as we will describe further below.

4 Approximations of the stochastic system

The master equation approach becomes for large system sizes N very time consuming, hence approximation schemes can help to speed up the analysis, which is especially important when analyzing empirical data typically with many stochastic runs for various parameter sets [1]. Here we apply the Kramers-Moya approximation obtaining a Fokker-Planck equation, from which we can sample individual realizations via a stochatic differential equation system [15, 16, 17], see especially for epidemiological systems like the ones treated here [27] for more details.

Fig. 5 shows the plots for the SIR systemcomparing the stochastic realization of the exact method from Gillespies algorithm for master equations in comparison with the faster Kramers-Moyal approximation leading to a Fokker-planck equation, which is the simulated via a stochastic differential equation system in Euler-Maruyama scheme. As well the over all dynamics as also the auto-correlations are well captured by the approximation.

5 The SIRI epidemic process

For a simple SIR-type model including reinfection, hence a simplest model in which primary and secondary infections can be destinguished, we have the following differential equation system

$$\begin{aligned} \frac{dS}{dt} &= \alpha R - \frac{\beta}{N}\, S \cdot (I + \varrho \cdot N) \\ \frac{dI}{dt} &= \frac{\beta}{N}\, S \cdot (I + \varrho \cdot N) + \theta \frac{\beta}{N}\, R \cdot (I + \varrho \cdot N) - \gamma I \\ \frac{dR}{dt} &= \gamma I - \alpha R - \theta \frac{\beta}{N}\, R \cdot (I + \varrho \cdot N) \end{aligned} \tag{8}$$

 ISBN: 978-84-616-9216-3

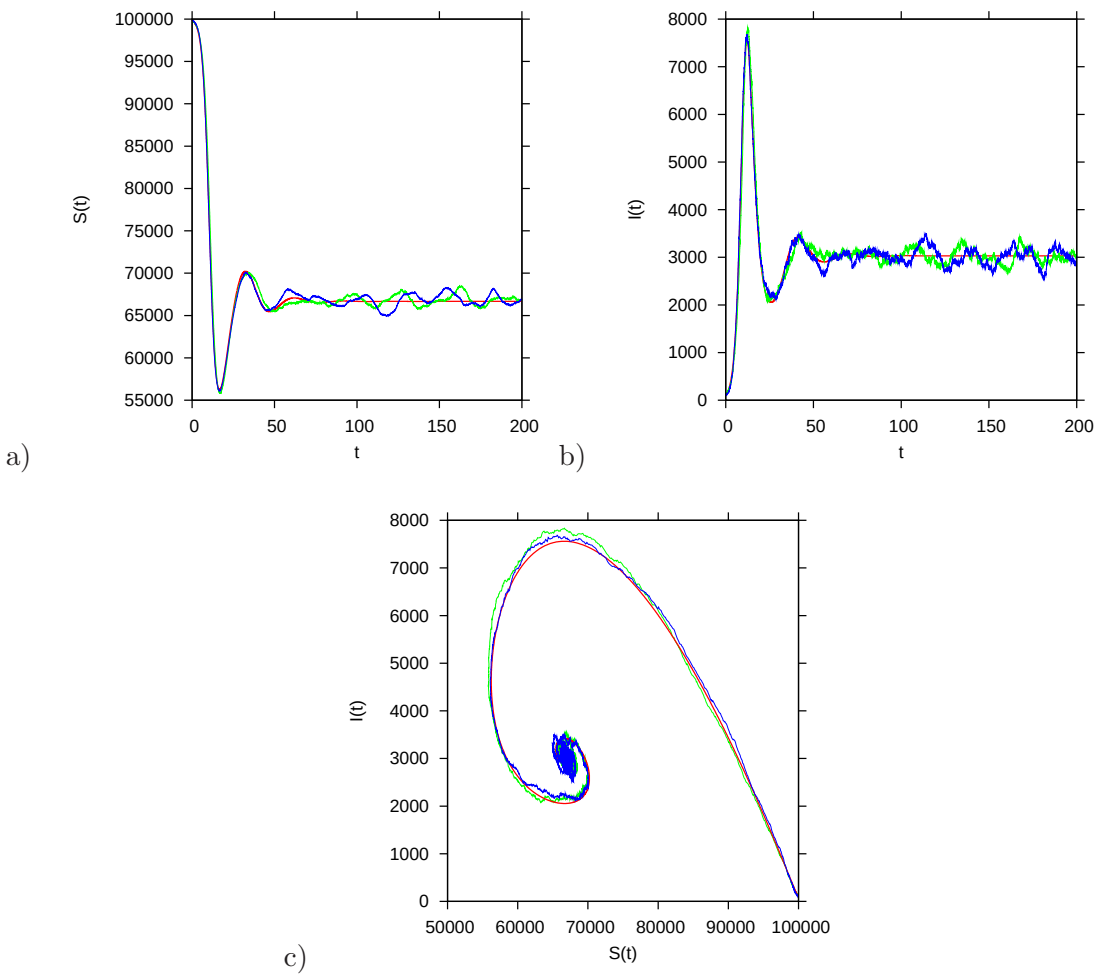


Figure 5: The stochastic SIR system in comparison of the exact Gillespie algorithm in green and the faster Fokker-Planck approximation. a) Time series of the susceptibles, b) of infected and c) in state space.

 ISBN: 978-84-616-9216-3

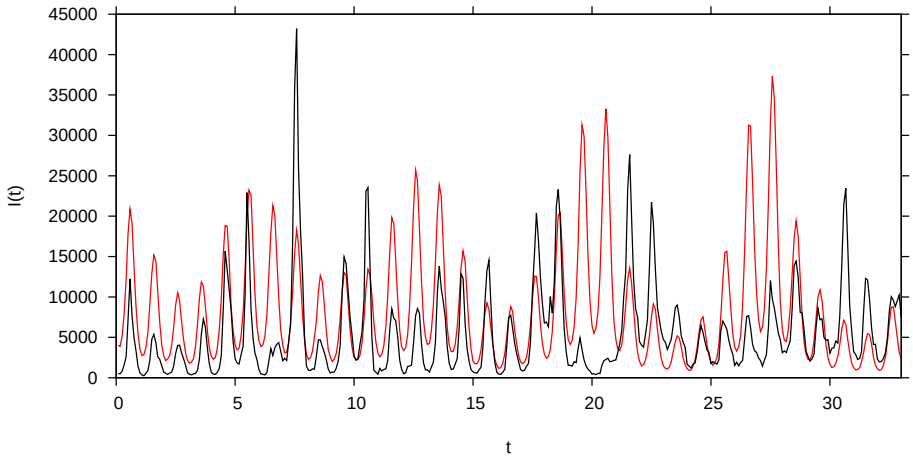


Figure 6: Comparison between the Thailand dengue incidence data set and the stochastic reinfection model SIRI simulation. Some of the qualitative features of the empirical data are already well represented by this seemingly quit oversimplified model, but go well along the experiences with Bayesian model comparison and its Occham's razor property quantified in the Bayes factor.

which describes the dynamics of the mean values for the total number of susceptibles, infected and recovered under the assumptions of mean field behaviour and homogeneous mixing as it would be obtained from a stochastic model, hence mean values of products can be replaced by products of means in the nonlinear contact term $(\beta/N)\ I \cdot S$ *etc.*

Due to the reinfection here we can in a simplest ansatz already include a distinction between primary and secondary infection, primary infection being the transition from S to I, and after recovery any secondary infection being the transition from R to I. The distinction between primary infection transitions and secondary ones can be easily performed in the Gillespie algorithm, as well as in the Euler-multinomial approximation. We also succeded in tracing this information easily in the Kramers-Moyal approximation. For a first model comparison between SIRI model output and the empirical data from severe dengue incidences in Thailand, see Fig. 6.

While the purely deterministic model shows only a fixed point, the stochastic model displays continued oscillations, of period length of roughly 7 years duration [27]. The model can further be forced, without loss of qualitative behaviour in the main frequencies [23], by a seasonal infection rate $\beta(t)$ and thus shows the already quite complex behaviour as visible in Fig. 6 and can roughly be compared with human severe dengue cases in Thailand. This simnplest reinfection model is somehow an oversimplified version of more realistic multi-strain models, as described in [4, 5, 6]. However, as a first model it can serve well, since often data cannot capture all complexity of a system and the simplest possible model gains most probability in a formal model comparison framework as provided e.g. by the Bayes

 ISBN: 978-84-616-9216-3

factor [2], an Occham's razor like feature. Only further statistical analyses along the lines outlined in [1] can eventually give further insigths into more complex dynamical behaviour as displayed in multi-strain models like deterministically chaotic attractors with positive Lyapunov exponents. We will now look at possible models for mosquito borne diseases with a similar Occhams razor approach, before more complex models can be treated in a similar rigor as the here presented. the staring point are models like the ones earlier analyzed by us [3] and [26].

6 The dynamics of mosquitos coupled with human infection models

The simplest model where the coupling between mosquito infection dynamics and human disease can be studied is the so called SISUV model [3] which after consideration of constant human population size and constant mosquito population size only have two coupled ordinary differential equations in the mean field case, or as stochastic model probability functions with only two variables. This case can be treated to quite some extend analystically and exhibits strong separation of time scales which can also be treated analytically by a center manifold analysis [3].

Here we present a stochastic system of SIRUV type, where the human disease model is an SIR model, which is more realistic for most vector borne diseases (the system already described and its deterministic version analyzed in detail in [3]). The disease vector part of mosquito infection is exactly like in the simpler SISUV model. Like in the above analyzed SIR system, also the SIRUV model shows deterministically a fixed point as attractor with oscillations into it. These oscillations are again revisited by the stochastic system, even when starting in the deterministic fixed point, see Fig. 7 for the enhanced stochastic oscillations in infected humans and in infected mosquitos.

In the next section we will describe how to combine the above mentioned models for human disease with reinfection [21, 22, 27] and the mosquito dynamics part from the previously investigated modquito models [3, 26].

7 The dynamics of mosquitos included in the reinfection model

Now we include the dynamics of susceptible mosquitos U and infected mosquitos V, which act as disease vectors, in the reinfection model for primary versus secondary infection in dengue fever, as it was describe before. We use essentially the notation for vector dynamics

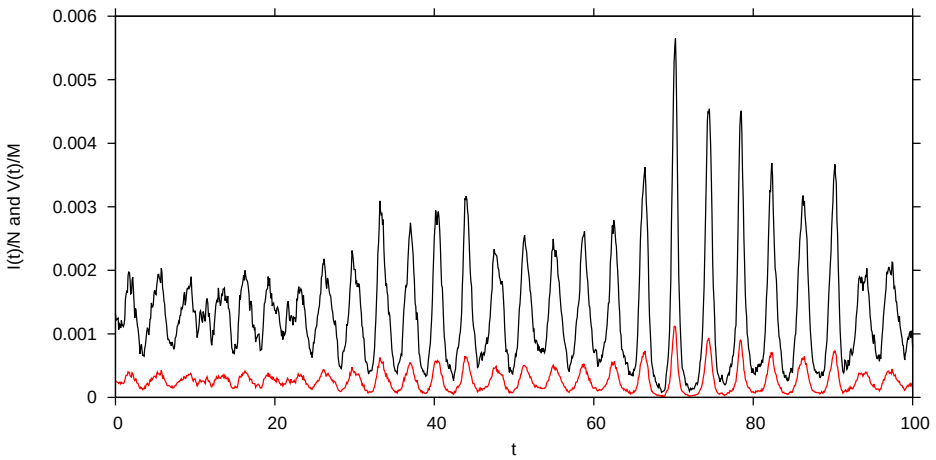


Figure 7: Time series of the stochastic SIRUV model for the densities of infected humans and infected mosquitos.

as described recently [3]. The model is given by the following differential equation system

$$\begin{aligned}
\frac{dS}{dt} &= \mu(N-S) - \frac{\beta}{M_0}\, S\cdot V - \varrho\beta S \\
\frac{dI}{dt} &= \frac{\beta}{M_0}\,(S+\theta R)\cdot V + \varrho\beta(S+\theta R) - (\gamma+\mu)I \\
\frac{dR}{dt} &= \gamma I - \mu R - \frac{\theta\beta}{M_0}\, RV - \varrho\theta\beta R \\
\frac{dU}{dt} &= \psi - \nu U - \frac{\vartheta}{N} UI \\
\frac{dV}{dt} &= \frac{\vartheta}{N} UI - \nu V
\end{aligned} \tag{9}$$

where eventually the birth rate of mosquitos can be seasonally forced via

$$\psi(t) = \psi_0\,(1+\psi_1 cos(\omega(t+\varphi))) \tag{10}$$

leading to time dependent total number of mosquitos $M(t) = U(t) + V(t)$ around a mean value of M_0. For further details on seasonal forcing in the mosquito dynamics see [26]. Under some conditions the dynamics of such systems can be simplified further by using a time scale separation argument, namely that the life time of humans and that of mosquitos in different in orders of magnitude, hence the mosquito infection dynamics is on a much faster time scale than the human disease dynamics [3].

 ISBN: 978-84-616-9216-3

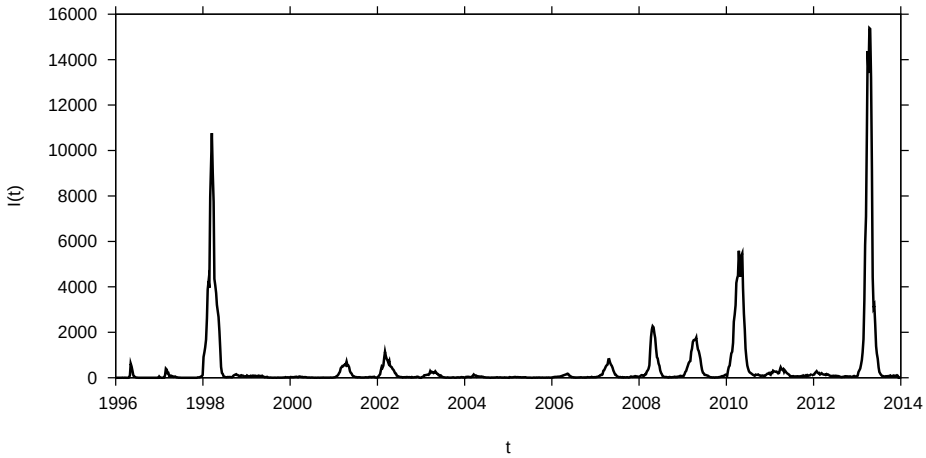


Figure 8: Dengue fever cases in the city of Belo Horizonte for 18 years recorded weekly.

8 Separation of time scales leads to simplified model

Taking into account that the mosquito dynamics is fast compared to the dynamics of human primary and secondary infection, hence

$$V(I(t)) = \frac{\frac{\vartheta}{\nu}\frac{I}{N}}{1+\frac{\vartheta}{\nu}\frac{I}{N}} \cdot M \tag{11}$$

we obtain the simplified model

$$\begin{aligned} \frac{dS}{dt} &= \mu(N-S) - \frac{\beta}{M_0}\, S \cdot \frac{\frac{\vartheta}{\nu}\frac{I}{N}}{1+\frac{\vartheta}{\nu}\frac{I}{N}} \cdot M - \varrho\beta S \\ \frac{dI}{dt} &= \frac{\beta}{M_0}\,(S+\theta R) \cdot \frac{\frac{\vartheta}{\nu}\frac{I}{N}}{1+\frac{\vartheta}{\nu}\frac{I}{N}} \cdot M + \varrho\beta(S+\theta R) - (\gamma+\mu)I \end{aligned} \tag{12}$$

with $R = N-S-I$ in a human population of constant size N, and in the non-forced system $M = M_0$, cancelling out the M in the ODE system. It is expected that this system also shows the stochastic amplification of transient oscillations of the deterministic model, as we observed in the original SIRUV model, as initially described in [3]. We will now give a first brief look into recently available data on mosquito abundance and the ratio of infected mosquitos versus overall numbers of mosquitos and human disease curves.

 ISBN: 978-84-616-9216-3

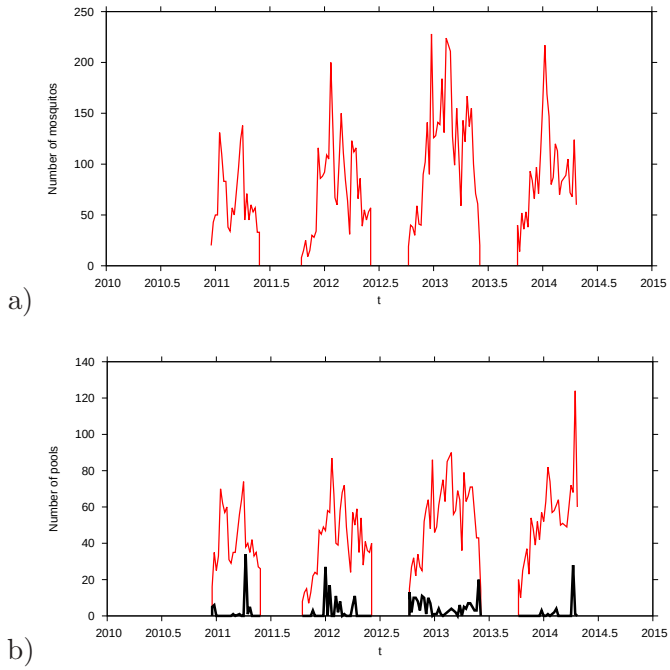


Figure 9: In the city of Belo Horizonte there were systematically mosquito traps placed since a few years to capture mosquitos transmitting dengue fever. a) Total number of captured mosquitos per week and b) total number of traps occupied (in red) and numbers of traps (pools) with infected mosquitos (in black).

9 Empirical data on dengue fever in humans and in mosquitos

In the city of Belo Horizonte in Minas Gerais in Brazil, we obtained human disease curves for dengue fever over a period of 18 years, see Fig. 8. There are also mosquito capturing experiments running, which will finish the 4th year collection of data by the end of May 2014, see Fig. 9. The first inspection of these data shows large fluctuations as well in the human disease curve, which vary from year to year significantly, as also in the overall number of mosquitos M, Fig. 9 a), and in the ratio of dengue virus infected mosquito pools and overall detected pools of mosquitos, Fig. 9 b), hence the ratio V/M. As opposed to other studies, here number of mosquitos in the pools are relatively low with many pools of only one mosquito. Earlier studies used much larger pool sizes due to financial restrictions of the dengue virus testing capacities.

The mosquito capturing study is still ongoing, and more results on the fluctuations in the mosquito dynamics versus the disease dynamics will be available only later. However, we can obtain first indications on basic features to be used in the modelling approach, such

 ISBN: 978-84-616-9216-3

as the ratio of infected to overall mosquitos. On a limited basis, also stochastic spatially extended models will be possible, since some of the data are collected with recording of collection locations, in ways of stochastic modelling indicated in [22] and recently [29]. Further approximation methods of the computationally demanding stochastic models can be applied as initially investigated in [28], see also [18, 19, 20].

Acknowledgements

This work has been supported by the European Union under FP7 in the project DENFREE and in various ways by FCT, Portugal, especially via the project PTDC/MAT/115168/2009. It also was supported by the following Brazilian agencies: Foundation of the Institute Oswaldo Cruz (FIOCRUZ), Brazilian Council for Scientific and Technological Development (CNPq), Thematic Programme of Support to Centers of Excellence (PRONEX Dengue Network) and Minas Gerais State Research Support Foundation (FAPEMIG). Juliana Bezerra is a doctoral student participating of the Sandwich Fellowship Program (CAPES) and by the Bill & Melinda Gates Foundation.

References

[1] Stollenwerk, N., Aguiar, M., Ballesteros, S., Boto, J., Kooi, B. & Mateus, L. (2012) Dynamic noise, chaos and parameter estimation in population biology, *Roy. Soc. Interface Focus* **2** 156–169.

[2] Mateus, L., Stollenwerk, N., & Zambrini, J.C. (2013) Stochastic Models in Population Biology: From Dynamic Noise to Bayesian Description and Model Comparison for Given Data Sets, *Int. Journal. Computer Math.* **90**, 2161–2173.

[3] Rocha, F., Aguiar, M., Souza, M., & Stollenwerk, N. (2013) Time-scale separation and center manifold analysis describing vector-borne disease dynamics, *Int. Journal. Computer Math.* **90**, 2105–2125.

[4] Aguiar, M., Ballesteros, S., Kooi, B.W., & Stollenwerk, N. (2011) The role of seasonality and import in a minimalistic multi-strain dengue model capturing differences between primary and secondary infections: complex dynamics and its implications for data analysis, *Jounal of Theoretical Biology*, **289**, 181–196.

[5] Aguiar, M., Stollenwerk, N. & Kooi, W. B. (2012). Scaling of stochasticity in dengue hemorrhagic fever epidemics. *Math. Model. Nat. Phenom.*, **7**, 1–11.

[6] Kooi, W. B., Aguiar, M., & Stollenwerk, N. (2013). Bifurcation analysis of a family of multi-strain epidemiology models, *Journal of Computational and Applied Mathematics*, **252**, 148–158.

[7] Aguiar, M., Kooi, W. B., Rocha, F., Ghaffari, P. & Stollenwerk, N. (2013). How much complexity is needed to describe the fluctuations observed in dengue hemorrhagic fever incidence data? *Ecological Complexity*, **16**, 31–40.

[8] Ionides, E.L., Breto, C., & King, A.A. (2006) Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA* **103**, 18438-18443.

[9] Stollenwerk, N., Drepper, F., & Siegel, H. (2001) Testing nonlinear stochastic models on phytoplankton biomass time series, *Ecological Modelling* **144**, 261–277.

[10] Stollenwerk, N., & Briggs, K.M. (2000) Master equation solution of a plant disease model, *Physics Letters* **A 274**, 84–91.

[11] Breto, C., He, D., Ionides, E.L., & King, A.A. (2009) Time series analysis via mechanistic models. *Annals of Applied Statistics* **3**, 319-348.

[12] Aguiar, M., Stollenwerk, N., & Kooi, B. (2009) Torus bifurcations, isolas and chaotic attractors in a simple dengue fever model with ADE and temporary cross immunity, *Intern. Journal of Computer Mathematics* **86**, 1867–77.

[13] Gillespie, D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22**, 403–434.

[14] Gillespie, D.T. (1978) Monte Carlo simulation of random walks with residence time dependent transition probability rates. *Journal of Computational Physics* **28**, 395–407.

[15] Honerkamp, J. (1993) *Stochastic Dynamical Systems: Concepts, Numerical Methods and Data Analysis*, VCH Publishers, Heidelberg, New York.

[16] van Kampen, N.G. (1992) *Stochastic Processes in Physics and Chemistry*, North-Holland, Amsterdam.

[17] Gardiner, C.W. (1985) *Handbook of stochastic methods* (Springer, New York).

[18] Gang, Hu (1987) Stationary solution of master equations in the large-system-size limit, *Physical Review* **A 36** 5782–5790.

[19] Elgart, V., & Kamenev, A. (2004) Rare event statistics in reaction-diffusion systems, arXiv:cond-mat/0404241v2, (also available as Physics Review E or similar).

 ISBN: 978-84-616-9216-3

[20] Billings, L., Mier-y-Teran-Romero, L., Lindley, B., & Schwartz, I. (2013) Intervention-based stochastic disease eradication, arXiv:1303.5614v1.

[21] Stollenwerk, N., van Noort, S., Martins, J., Aguiar, M., Hilker, F., Pinto, A. & Gomes, G. (2010) A spatially stochastic epidemic model with partial immunization shows in mean field approximation the reinfection threshold, *Journal of Biological Dynamics* **4**, 634–649.

[22] N. Stollenwerk and V. Jansen, *Population biology and criticality*, Imperial College Press, London, 2011.

[23] R.P. Boland, T. Galla, A.J. McKane, Limit cycles, complex Floquet multipliers and intrinsic noise, *arXiv:0903.5248v1.*

[24] A.J. McKane, T.J. Newman, Predator-prey cycls from resonant amplification of demographic stochasticity, *Phys. Rev. Lett.* **94** (2005) 218102–7.

[25] D. Alonso, A. McKane, M. Pascual, Stochastic Amplification in Epidemics, *Journal of the Royal Society Interface*, **4**, (2006)575–582.

[26] Rocha, F., Skwara, U., Aguiar, M., Stollenwerk, N. (2013). Understanding dengue fever dynamics: study of seasonality in the models. *Proceedings of the 13th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2013*, pp 1197-1209 ISBN: 978-84-616-2723-3, edited by Jesus Vigo et al., Almeria.

[27] Nico Stollenwerk, Maira Aguiar, Filipe Rocha, Urszula Skwara. (2013). Testing particle filters for dengue fever studies via simple reinfection models. *Proceedings of the 13th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2013*, pp 1262-1277 ISBN: 978-84-616-2723-3, edited by Jesus Vigo et al., Almeria.

[28] Nico Stollenwerk, Davide Masoero, Urszula Skwara, Filipe Rocha, Peyman Ghaffari, Maira Aguiar. (2013). Semiclassical approximations of stochastic epidemiological processes towards parameter estimation. *Proceedings of the 13th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2013*, pp 1278-1289 ISBN: 978-84-616-2723-3, edited by Jesus Vigo et al., Almeria.

[29] Urszula Skwara, Filipe Rocha, Maira Aguiar, Nico Stollenwerk. (2013). Superdiffusion in epidemiological models. *Proceedings of the 13th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2013*, pp 1250-1261 ISBN: 978-84-616-2723-3, edited by Jesus Vigo et al., Almeria.

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

The role of stiffness in the proliferation of brain tumors

J. R. Branco[1], J. A. Ferreira[2] and Paula de Oliveira[2]

[1] *CMUC & Polytechnic Institute of Coimbra, ISEC, DFM, Coimbra, Portugal*

[2] *CMUC & Department of Mathematics, University of Coimbra, Coimbra, Portugal*

emails: jrbranco@isec.pt, ferreira@mat.uc.pt, poliveir@mat.uc.pt

Abstract

In this paper we present a mathematical model to describe the evolution of glioma cells taking into account the viscoelastic properties of brain tissue. A theoretical stability analysis gives information to design protocols which efficiency is illustrated by a number of numerical simulations.

Key words: Glioma, viscoelastic behaviour, chemotherapy, numerical simulation.

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. Extensive research has been done to model cancerous growth, however the understanding of malignant gliomas is much less complete, mostly because migration of gliomas represent a very challenging problem from a mathematical viewpoint.

Gliomas are diffusive and highly invasive brain tumors. 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 from the high mobility of glioma cells, which is partly driven by the mechanical properties of brain tissue.

The first model to measure the growth of an infiltrating glioma was provided by Murray in the early 90s ([19]). He formulated the problem as a conservation law where the rate of change of tumor cell population results from mobility and net proliferation of cells. An equation of type

$$\frac{\partial c}{\partial t} = \nabla.(\tilde{D}\,\nabla c) + f(c) \ \text{ in } \Omega \times (0,\infty) \tag{1}$$

 ISBN: 978-84-616-9216-3

was used, where $\Omega \subset \mathbb{R}^n, n = 1, 2, 3$, is the glioma domain, $c(x, t)$ denotes the tumor cell density at location x and time t, $f(c)$ denotes net proliferation of tumor cells (generally assumed to be exponential, $f(c) = \rho\, c$ where the net proliferation rate ρ is constant), $\tilde{D}$ is the diffusion tensor and ∇ defines the spatial gradient operator.

The partial differential equation (1), of parabolic type, was established combing the mass conservation law with Fick's law for the mass flux J_F,

$$J_F = -\tilde{D}\,\nabla c\,. \tag{2}$$

It is well known that that Fickian approach gives rise to infinite speed of propagation which is not physically observable. To avoid the limitation of Fickian models an hyperbolic correction has been proposed in different contexts (see [1], [6], [9], [10], [15], [17], and [20]).

The aim of this paper is to establish a class of non Fickian models that take into account the viscoelastic behavior of the brain tissue and to present a stable numerical method for this class of models. A simplified version of this model was considered [2] using a simple geometry. To apply the modeling approach to specific patients a more realistic look at the brain geometry and structure is necessary. In this case we can follow [23] where a complex geometry of the brain and a space dependent diffusion coefficient were considered to reflect the observation that glioma cells exhibit higher motility in the white matter than in grey matter ([14]).

We observe that the most popular treatments used to combat gliomas are chemotherapy and radiotherapy. Chemotherapy involves the use of drugs to disrupt the cell cycle and to block proliferation. Tracqui *et al.* [24] incorporated chemotherapy by introducing cell death as a loss term. If $G(t)$ defines the rate of cells death then, assuming a loss proportional to the tumour cells density, equation (1) is replaced by

$$\frac{\partial c}{\partial t} = \nabla.(\tilde{D}\nabla c) + f(c) - G(t)c \ \text{ in } \Omega \times (0, T]\,, \tag{3}$$

where

$$G(t) = \begin{cases} k, & \text{when chemotherapy is being administered} \\ 0, & \text{otherwise}\,. \end{cases} \tag{4}$$

Here k describes the rate of cell death due to exposure to the drug. The main question is how to define k and the periods of chemotherapy applications that lead to control the glioma mass.

2 A viscoelastic model

The brain tissue presents a viscoelastic behaviour that can be described by the Voigt-Kelvin model ([13], [16], [18]). In this section we present a class of non Fickian models to describe

 ISBN: 978-84-616-9216-3

the space and time evolution of glioma cancer cells, combining the diffusion process with the viscoelastic properties of the brain tissue.

Several authors have studied the diffusion in a viscoelastic medium ([5], [7], [8] and [22]), using a modified diffusion equation of type

$$\frac{\partial c}{\partial t} = \nabla.(\tilde{D}\nabla c) + \nabla.(\tilde{D}_v\nabla\sigma) + f(c) \text{ in } \Omega \times (0,\infty), \tag{5}$$

where σ represents the stress exerted by the medium on the diffusing molecules and $\tilde{D}$ represents a diagonal tensor with positive entries.

Even if studies of glioma growth have essentially addressed biochemical and genetic factors, recent biomedical research has highlighted the role of mechanical properties. Our aim in this paper is the modelling and analysis of glioma growth under the effect of the rheological properties of the brain tissue.

Investigators have observed that the stiffness of extracellular matrix can either increase or decrease the diffusion of migration cells. These observations are explained by the fact that extracellular matrix stiffness induce complex biochemical phenomena that depend on the type of diffusive cells and microenvironment properties.

In [25] the authors observed in vitro migration of fibroblasts from soft to stiff regions of extracellular matrix. Following this paper we consider equation (5) where $\tilde{D}_v$ is a diagonal tensor with negative entries.

We assume that the viscoelastic behaviour of the brain tissue is described by the Voigt-Kelvin model

$$\frac{\partial \sigma}{\partial t} + \beta\sigma = \alpha_1\epsilon + \alpha_2\frac{\partial \epsilon}{\partial t}\,, \tag{6}$$

where ϵ stands for the strain. Equation (6) is based on a mechanistic model which is represented by a spring and a dashpot in parallel, connected with a free spring. In (6) the viscoelastic characteristic time β is given by $\beta = \frac{E_0+E_1}{\mu_1}$, and $\alpha_1 = \frac{E_0E_1}{\mu_1}$, $\alpha_2 = E_0$ where E_1 is the Young modulus of the spring element, μ_1 represents the viscosity and E_0 stands for the Young modulus of the free spring (see [13], [16], [18]).

If we assume that the strain ϵ satisfies $\epsilon = \lambda c$ where λ is a positive constant (see [5], [7] and [8]), from (6) we obtain

$$\frac{\partial c}{\partial t} = \nabla.(D\nabla c) + \int_0^t k_{er}(t-s)\nabla.(D_v\nabla c(s))\,ds + f(c) \text{ in } \Omega \times (0,\infty)\,, \tag{7}$$

where $D = \tilde{D} + \lambda\alpha_2\tilde{D}_v$, $D_v = \lambda(\alpha_1 - \beta\alpha_2)\tilde{D}_v$ and $k_{er}(s) = e^{-\beta s}$.

According to [11] and [12] we will consider the following assumptions: glioma cells are of two phenotypes - proliferation (state 1) and migratory (state 2); in state 2 cells randomly move but there is no cell fission; in state 1 cancer cells do not migrate and only proliferation takes place with rate ρ; a cell of type 1 remains in state 1 during a time period and then

 ISBN: 978-84-616-9216-3

switches to a cell of type 2; β_1 is the switching rate from state 1 to 2; a cell of type 2 remains in state 2 during a time period and then switches to a cell of type 1; β_2 is the switching rate from state 2 to 1.

Let $u(x,t)$ and $v(x,t)$ represent the density of migratory and proliferation cells at x and t, respectively. The dynamics of glioma cells is then described by

$$\begin{cases} \dfrac{\partial u}{\partial t} &= \nabla.(D\,\nabla u) + \displaystyle\int_0^t k_{er}(t-s)\,\nabla.(D_v \nabla u(s))\,ds - \beta_1 u + \beta_2 v \ \text{ in } \Omega \times (0,T], \\ \dfrac{\partial v}{\partial t} &= \rho v + \beta_1 u - \beta_2 v \ \text{ in } \Omega \times (0,T], \end{cases} \tag{8}$$

where D and D_v denote square matrices of order n. The set of equations (8) is complemented with initial conditions

$$u(0) = u_0, \ \ v(0) = v_0 \ \text{ in } \ \Omega\,,$$

where u_0 and v_0 define the initial spatial distribution of malignant cells, and boundary conditions

$$J.\eta = 0 \ \ \text{ on } \partial\Omega, \tag{9}$$

where $\partial\Omega$ denotes the boundary of Ω, η represents the exterior unit normal to the brain region and the non Fickian flux J is given by $J(t) = -D\nabla u(t) - \displaystyle\int_0^t e^{-\beta(t-s)} D_v \nabla u(s)\,ds\,.$ Condition (9) means that the glioma is located inside of the brain and the cancer cells do not cross the pia mater.

We will assume that $D = [d_{ij}]$ and $D_v = [d_{v,ij}]$ are diagonal matrices with diagonal entries d_i and $d_{v,i}$ such that

$$0 < d_i,\, d_{v,i} \ \text{ in } \ \ \overline{\Omega},\, i = 1, \ldots, n. \tag{10}$$

If we consider the mass of glioma cells in Ω, $\ M_1(t) = \displaystyle\int_\Omega (u(t) + v(t))\,dx$ we showed in [4] that $M_1(t) \le e^{\rho t} M_1(0)$, assuming the positivity of u, which means that mass $M_1(t)$ of cancer cells at time t depends on the initial mass, on time t and on the proliferation rate ρ.

To avoid the positivity assumption on u we consider the mass related functional $M_2(t) = \|u(t)\|^2 + \|v(t)\|^2$, where $\|.\|$ denotes the usual L^2. In this case we deduce that

$$M_2(t) \le e^{2\max\{\frac{\beta_2-\beta_1}{2}, \frac{\beta_1-\beta_2}{2}+\rho, -\beta\}t} M_2(0)\,. \tag{11}$$

If the tumor density is largen than 1 then an upper bound for $M_1(t)$ can be deduced from an estimate of $M_2(t)$. We observe that we can not select parameters β_1, β_2, ρ such that $M_2(t)$ is bounded in time. We also remark that inequality (11) allow us to conclude the stability of the proposed mathematical model with respect to perturbations of the initial conditions in $[0,T]$, for fixed $T > 0$.

 ISBN: 978-84-616-9216-3

3 Chemotherapy: control of the glioma growth

In this section we study the behaviour of the glioma mass when chemotherapy is considered and we establish criteria to define protocols that lead to the decreasing of the tumor mass. All the results of this section were carefully analyzed in [3].

To take into account the chemotherapy effect, the viscoelastic model for glioma growth (8) is modified as follows

$$\begin{cases} \dfrac{\partial u}{\partial t} &= \nabla.(D\,\nabla u) + \displaystyle\int_0^t k_{er}(t-s)\,\nabla.(D_v \nabla u(s))\,ds - \beta_1 u + \beta_2 v - G(t)u \ \text{ in } \Omega\times(0,T], \\ \dfrac{\partial v}{\partial t} &= \rho v + \beta_1 u - \beta_2 v - G(t)v \ \text{ in } \Omega\times(0,T], \end{cases} \tag{12}$$

where $G(t)$ is defined by (4).

Considering $E(t) = M_2(t) + \| \displaystyle\int_0^t k_{er}(t-s)\sqrt{D_v}\nabla u(s)\,ds\|^2$, it can be proved that

$$E'(t) \le 2\max\Big\{\frac{\beta_2-\beta_1}{2} - G(t), \frac{\beta_1-\beta_2}{2} + \rho - G(t), -\beta\Big\}E(t)\,. \tag{13}$$

From (13) some conditions on the parameters, that lead to a decreasing of $M_2(t)$, can be established:

1. If the net proliferation rate is greater than the switching proliferation rate

$$\rho > \beta_2 - \beta_1\,, \tag{14}$$

and the total amount of death cells until time t due to chemotherapy effect is such that

$$\Big(\frac{\beta_1-\beta_2}{2} + \rho\Big)t < \int_0^t G(s)\,ds < \big(\frac{\beta_2-\beta_1}{2} + \beta\big)\,t\,, \tag{15}$$

then we can conclude that $M_2(t)$ decreases.

From (15) we conclude that the difference between the net and switching proliferation rates should be less than the viscoelastic characteristic time, that is,

$$\rho - (\beta_2 - \beta_1) < \beta. \tag{16}$$

If no viscoelastic effects are considered ($\beta = 0$) we deduce from (15) that $\displaystyle\int_0^t G(s)\,ds$, which measures in some sese the intensity of the treatment, should be smaller.

 ISBN: 978-84-616-9216-3

2. Otherwise, if the net proliferation rate is less than the switching proliferation rate

$$\rho < \beta_2 - \beta_1 \tag{17}$$

and the total amount of death cells until time t, due to chemotherapy effect, is such that

$$\Big(\frac{\beta_2 - \beta_1}{2}\Big)t \;<\; \int_0^t G(s)\,ds \;<\; \Big(\frac{\beta_1 - \beta_2}{2} + \rho + \beta\Big)t\,, \tag{18}$$

then we conclude that $M_2(t)$ decreases. Again we observe that the parameter β has influence on the admissible threshold of the chemotherapy treatment.

We note that condition (18) implies

$$\rho - (\beta_2 - \beta_1) > \beta\,. \tag{19}$$

When chemotherapy is applied, conditions (15) and (18) can be used to determine an effective dosage that induces a rate k of cell death due to the exposure to the drug that allows to control the total tumor mass. Obviously the value of k depends of the protocol of chemotherapy. The typical bang-bang protocol corresponds to treatment which alternate maximum doses of chemotherapy with rest periods when no drug is administered, as defined by (4) and illustrated in Figure 1.

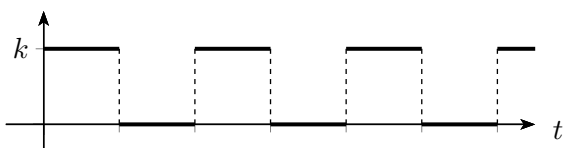


Figure 1: Chemotherapy protocol.

4 A fully discrete model

In this section we present a stable method to obtain numerical approximations for the density of proliferation and migratory glioma cells. We show that the method preserves the qualitative behaviour of the initial boundary value problem studied in the last section.

We assume that $n = 2$, Ω is the square $[0, L] \times [0, L]$ and $H = (h_1, h_2)$ with $h_i > 0, i = 1, 2$. In $\overline{\Omega}$ we introduce the spatial grid $\overline{\Omega}_H = \{(x_{1,i}, x_{2,j}), i = 0, \ldots, N_{h_1}, j = 0, \ldots, N_{h_2}\}$,where $x_{\ell,i} = x_{\ell,i-1} + h_\ell$, $i = 1, \ldots, N_{h_\ell}$, $x_{\ell,0} = 0$, $x_{\ell,N_{h_\ell}} = L$, for $\ell = 1, 2$. By $\partial\Omega_H$ we represent the set of boundary points. We introduce the following auxiliary points $x_{\ell,-1} = x_{\ell,0} - h_\ell$, $x_{\ell,N_{h_\ell}+1} = x_{\ell,N_{h_\ell}} + h_\ell, \ell = 1, 2$.

 ISBN: 978-84-616-9216-3

Let $w_H = (u_h,\, v_H)$ represent a semi-discrete aprocimation of $w = (u, v)$. To simplify the presentation we use the notation $w_{i,j} = w_H(x_{1,i}, x_{2,j})$. We discretize $\dfrac{\partial}{\partial x_1}(a\dfrac{\partial u}{\partial x_1})$, a is a scalar functions, using the usual second order finite difference discretization

$$\nabla^*_{h_1}(\hat{a}_H \nabla_{h_1} u_H)(x_{1,i}, x_{2,j}) = \frac{1}{h_1}\Big(a_{i+1/2,j} D_{-x_1} u_{i+1,j} - a_{i-1/2,j} D_{-x_1} u_{i,j}\Big), \tag{20}$$

where $a_{i\pm 1/2,j} = a(x_{1,i} \pm \frac{h_1}{2}, x_{2,j})$ and D_{-x_1} denotes the usual backward finite difference operator in x_1 direction. The second order finite difference discretization $\nabla^*_{h_2}(\hat{b}_H \nabla_{h_2} u_H)(x_{1,i}, x_{2,j})$ to discretize $\dfrac{\partial}{\partial x_2}(b\dfrac{\partial u}{\partial x_2})$ is defined analogously.

In $[0, T]$ we introduce the grid $\{t_n, n = 0, \ldots, M\}$ with $t_n = t_{n-1} + \Delta t, n = 1, \ldots, M$, $t_0 = 0, t_M = T$. To compute numerical approximations for u and v in $(x_{1,i}, x_{2,j})$ at time level t_n, $u^n_H(x_{1,i}, x_{2,j})$, $v^n_H(x_{1,i}, x_{2,j})$, respectively, we introduce the fully discrete initial boundary value problem

$$\begin{cases} D_{-t} u^{n+1}_H = \displaystyle\sum_{i=1,2} \nabla^*_{h_i}(d_i \nabla_{h_i} u^{n+1}_H) + \Delta t \sum_{\ell=1}^{n+1} k_{er}(t_{n+1} - t_\ell) \sum_{i=1,2} \nabla^*_{h_i}(d_{v,i} \nabla_{h_i} u^\ell_H) \\ \qquad\qquad -(\beta_1 + G(t_{n+1}) u^{n+1}_H + \beta_2 v^{n+1}_H \ \text{ in } \overline{\Omega}_H, \\ D_{-t} v^{n+1}_H = (\rho - \beta_2 - G(t_{n+1})) v^{n+1}_H + \beta_1 u^{n+1}_H \ \text{ in } \overline{\Omega}_H, \\ \qquad\qquad n = 0, \ldots, M-1, \end{cases} \tag{21}$$

$$u^0_H = u_0, \ \ v^0_H = v_0 \ \text{ in } \overline{\Omega}_H, \tag{22}$$

$$\begin{aligned} D_{\eta_{x_1}} u^{n+1}_H(x_{1,i}, x_{2,j}) = 0, \ i = 0, N_{h_1}, \ j = 0, \ldots, N_{h_2}, \\ D_{\eta_{x_2}} u^{n+1}_H(x_{1,i}, x_{2,j}) = 0, \ i = 0 \ldots, N_{h_1}, \ j = 0, N_{h_2}\,, \end{aligned} \tag{23}$$

where

$$D_{\eta_{x_1}} u^{n+1}_H(x_{1,i}; x_{2,j}) = D_{d_1, \eta_{x_1}} u_H(x_{1,i}; x_{2,j}) + \Delta t \sum_{l=1}^{n+1} k_{er}(t_{n+1} - t_l) D_{d_{v,1}, \eta_{x_1}} u^l_H(x_{1,i}; x_{2,j})\,, \tag{24}$$

and $D_{a,\eta_{x_1}} u_H(x_{1,i}; x_{2,j})$ is defined by

$$\frac{1}{2}\Big(a(x_{1,i+1/2}; x_{2,j}) D_{-x_1} u^{n+1}_H(x_{1,i}; x_{2,j}) + a(x_{1,i-1/2}; x_{2,j}) D_{-x_1} u^{n+1}_H(x_{1,i}; x_{2,j})\Big)\,,$$

for $a = d_1,\, d_{v,1}\,,$ being $D_{a,\eta_{x_2}} u_H(x_{1,i}; x_{2,j})$ defined analogously.

We now study the stability of the discrete scheme (21), (22) and (23). It's easy to prove that

$$\min\{1, 1 - \Delta t \alpha_{n+1}\}\, E^{n+1}_H \le E^n_H, \ n = 0, \ldots, M, \tag{25}$$

 ISBN: 978-84-616-9216-3

where

$$E_H^n = M_H^n + \sum_{i=1,2} \|\Delta t \sum_{\ell=0}^{n} k_{er}(t_n - t_\ell)\sqrt{\hat{d}_{v,i,H}} D_{-x_i} u_H^\ell\|_{h_i}^2,$$

$M_H^n = \|u_H^n\|_H^2 + \|v_H^n\|_H^2$ represents a discretization of $M_2(t)$ and

$$\alpha_n = 2\Delta t \max\big\{\frac{\beta_2 - \beta_1}{2} - G(t_n), \rho + \frac{\beta_1 - \beta_2}{2} - G(t_n)\big\}.$$

From (25) we deduce the stability inequality

$$E_H^{n+1} \leq \prod_{\ell=1}^{n+1} \frac{1}{\min\{1, 1 - \alpha_\ell \Delta t\}} E_H^0\,, \tag{26}$$

provided that

$$1 - \Delta t\, \alpha_\ell \,>\, 0\,,\ \text{for all}\ \ell\,. \tag{27}$$

When G is defined by (4), if the administered dosage of drug is fixed such that

$$\frac{\beta_2 - \beta_1}{2} \,>\, k, \quad \rho + \frac{\beta_1 - \beta_2}{2} \,>\, k\,, \tag{28}$$

then condition (27) holds provided that time step size Δt satisfies

$$\Delta t < \frac{1}{\alpha_\beta}, \tag{29}$$

where

$$\alpha_\beta = 2\max\big\{\frac{\beta_2 - \beta_1}{2}, \frac{\beta_1 - \beta_2}{2} + \rho\big\}.$$

In this case (26) can be rewritten as follows

$$E_H^{n+1} \leq \frac{1}{\Big(1 - 2\Delta t \alpha_\beta\Big)^{(n+1)}}\, E_H^0\,,$$

and consequently

$$E_H^{n+1} \leq e^{\frac{2(n+1)\Delta t}{1 - 2\Delta t \alpha_\beta}}\, E_H^0\,, \tag{30}$$

which means that the numerical scheme (21), (22), (23) is conditionally stable under the condition (29) provided that the coefficients $\beta_i, i = 1, 2$, and ρ satisfy (28).

 ISBN: 978-84-616-9216-3

5 Numerical results

In this section we illustrate the behaviour of (21), (22) and (23). We consider a homogeneous square domain $\Omega = [0, 15\,cm] \times [0, 15\,cm]$, growth rate $\rho = 0.012\,/day$ and switching parameters $\beta_1 = 10^{-6}/day$ and $\beta_2 = 0.036/day$. These values are physiological and have been obtained from [21]. According to [18] the initial condition is defined by $10^5\,cells/cm^2$ proliferation tumor cells located at the middle point of the domain, $E_0 = 3156\,Pa$, $E_1 = 6E_0$ and $\mu = 8.9{\times}10^{-4}\,Pa{\cdot}s$. We also consider an isotropic behaviour with $\tilde{d}_{11} = \tilde{d}_{22} = 0.004\,cm^2/day$ and $\tilde{d}_{v,11} = \tilde{d}_{v,22} = -10^{-14}\,/Pa \cdot day$ (which leads to $d_{11} = d_{22} \sim 0.004\,cm^2/day$ and $d_{v,11} = d_{v,22} = 0.001\,cm^2/day^2$) and parameter $\lambda = 1\,cm^2$.

Let us consider that the chemotherapy treatment is defined by (4) and applied with a protocol as illustrated in Figure 1. Conditions (15) are used to compute a profile for $G(t)$ that lead to control the total tumor mass. We consider a 24h dosage and different rest periods. In Table 1 we show the minimum value of k.

Protocol	$k_{\min}\,[./day]$
each 7 days	0.224
each 14 days	0.448

Table 1: $k_{\min}$ as (15), for a protocol of 24 consecutive hours of chemotherapy.

In Figure 2 we compare glioma masses for tree patients: one untreated and two submitted to chemotherapy starting at day 7 and with 7 and 14 rest periods, respectively. The values of k were computed using conditions (15). We observe a significant reduction of glioma masses when compared to glioma's untreated patient. The results presented in this figure show the effectiveness of our approach to define chemotherapy protocols.

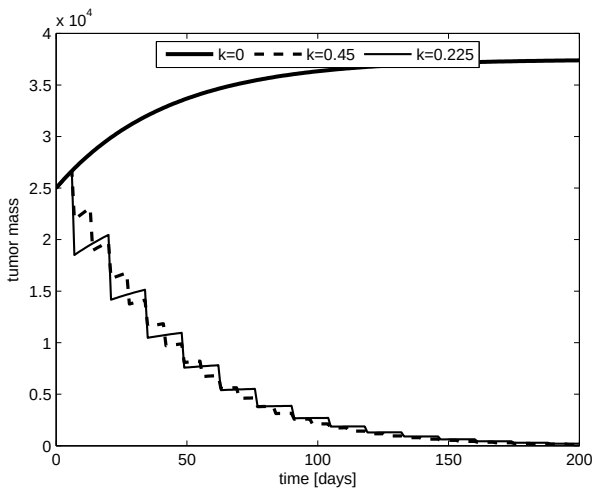


Figure 2: Glioma masses $M_1(t)$ for 200 days.

In Figure 3 we plot the numerical solutions at day 104 for $E_0 = 3156\,Pa$ Solutions are presented in a logarithmic scale, which means that the contour plots represent the power of 10 of the density of tumor cells. For both cases we also present the distribution of proliferation cells for two patients submitted at chemotherapy protocol with a 24h dosage and 14 days of rest period (dosage at days 7, 21, 35, 49, etc). Values of k were computed using conditions (15) according to the weaker restriction. We observe a more intensive spreading when Young modulus (of the free spring) increases. This conclusion is in agreement with experimental results as stated in [25].

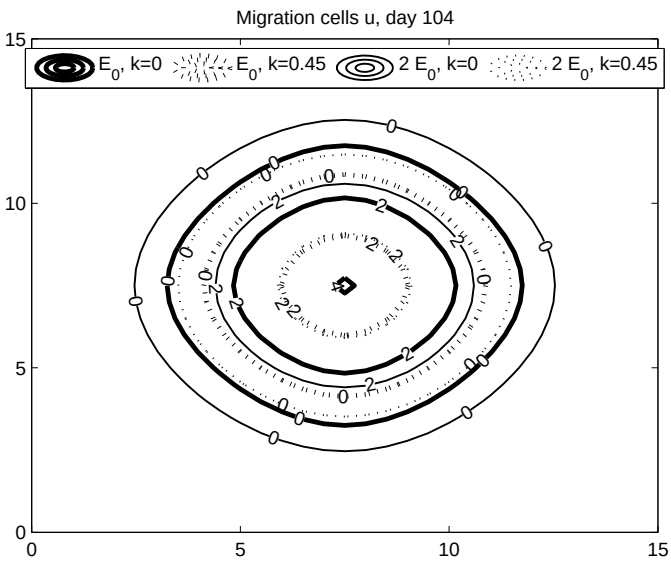


Figure 3: Distribution of proliferation cells results at day 104 ($E_0 = 3156\,Pa$).

6 Conclusions

In this paper we studied a mathematical model to describe the evolution of glioma cells with and without chemotherapy. The model was established combining a mass conservation law with a non Fickian mass flux that takes into account the viscoelastic behaviour of the brain tissue described by the Voigt-Kelvin model.

We deduced estimates that allowed to define sufficient conditions on the parameters that lead to control the glioma mass.

A fully discrete scheme was defined and the stability of such scheme was analyzed.

Numerical experiments suggest that our approach is a promising one. The behaviour of the mass of glioma cells was illustrated under the conditions deduced for the chemotherapy protocols.

 ISBN: 978-84-616-9216-3

Acknowledgements

This work was partially supported by the Centro de de Matemática da Universidade de Coimbra (CMUC), funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e Tecnologia under the projects PEst-C/MAT/UI0324/2011.

References

[1] J. R. Branco, J. A. Ferreira, P. de Oliveira, *Numerical methods for the generalized Fisher-Kolmogorov-Petrovskii-Piskunov equation,* Appl. Numer. Math., **57** (2007) 89–102.

[2] J. R. Branco, J. A. Ferreira, P. de Oliveira, *Memory in mathematical modelling oh highly diffusive tumors,* in 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 , (2011) 242–253.

[3] J. R. Branco, J. A. Ferreira, P. de Oliveira, *Efficient protocols to control glioma growth,* in Proceedings of the 2013 International Conference on Computational and mathematical Methods in Science and Engineering, Almeria, Spain, June 24-27, 2013, Editors: I. Hamilton and J. Vigo-Aguiar, (2013) 293–303.

[4] J. R. Branco, J. A. Ferreira, P. de Oliveira, *A viscoelastic model for glioma growth,* in Proceedings of SIMULTECH 2013 - 3rd International Conference on Simulation and Modeling, Technologies and Applications, Methodologies, Reykjavik, Iceland, 29-31 July, 2013, Editors: Tuncer Ãren, Janusz Kacprzyk, Leifur Leifsson, Mohammad S. Obaidat and Slawomir Koziel, (2013) 689–695.

[5] D. A. Edwards, D. S. Cohen, *An unusual moving boundary condition arising in anomalous diffusion problems,* SIAM J. Appl. Math. **55** (1995) 662–676.

[6] D. A. Edwards, D. S. Cohen, *A mathematical model for a dissolving polymer,* AIChE J. **41** (1995) 2345–2355.

[7] D. A. Edwards, *Non-Fickian diffusion in thin polymer films,* Poplym. Sci. Ser. B+ **34** (1996) 981–997.

[8] D. A. Edwards, *A spatially nonlocal model for polymer-penetrant diffusion,* J. Appl. Math. and Phys., **52** (2001) 254–288.

[9] S. Fedotov, *Traveling waves in a reaction-diffusion system: diffusion with finite velocity and Kolmogorov-Petrovskii-Piskunov kinetics,* Phys. Rev. E **58** (1998) 4:5143–5145.

[10] S. Fedotov, *Nonuniform reaction rate distribution for the generalized Fisher equation: Ignition ahead of the reaction front,* Phys. Rev. E **60** (1999) 4:4958–4961.

[11] S. Fedotov, A. Iomin, *Migration and proliferation dichotomy in tumor-cell invasion,* Phys. Rev. Lett. **98** (2007) 118110(1)–(4).

[12] S. Fedotov, A. Iomin, *Probabilistic approach to a proliferation and migration dichotomy in tumor cell invasion,* Phys. Rev. E **77** (2008) 1031911(1)–(10).

[13] G. Franceschini, *The mechanics of human brain tissue,* PhD thesis, University of Trento, 2006.

[14] A. Giese, L. Kluwe, B. Laube, H. Meissner, M. Berens, M. Westphal, *Migration of human glioma cells on myelin,* Neurosurgery **38** (1996) 755–764.

[15] S. Hassanizadeh, *On the transient non-Fickian dispersion theory,* Transp. Porous Media **23** (1996) 107–124.

[16] J. Humphrey, *Continuum biomechanics of soft biological tissues,* Proceedings of Royal Society London **459** (2003) 3–46.

[17] D. Joseph, L. Preziosi, *Heat waves,* Rev. Mod. Phys. **61** (1989) 47–71.

[18] A. Mehrabian, Y. Abousleiman, *General solutions to poroviscoelastic model of hydrocephalic human brain tissue,* J. Theor. Biol. **291** (2011) 105–118.

[19] J. D. Murray, *Mathematical Biology,* Springer, 2002.

[20] S. P. Neuman, D. M. Tartakovsky, *Perspective on theories of anomalous transport in heterogeneous media,* Adv. Water Resour. **32** (2009) 670–680.

[21] A. Roniotis, K. Marias, V. Sakkalis, M. Zervakis, *Diffusive modelling of glioma evolution: a review,* J. Biomed. Eng., **3** (2010) 501-508.

[22] S. Shaw, J. R. Whiteman, *Some partial differential Volterra equation problems arising in viscoelasticity,* Proceeding of the Conference on Differential Equations and their Applications, Brno, (1997) 183–200.

[23] K. R. Swanson, E. C. Alvord Jr, J. D. Murray, *A quantitative model for differential motility of gliomas in grey and white matter,* Cell Proliferat. **33** (2000) 317–329.

[24] P. Tracqui, G. C. Cruywagen, D. E. Woodward, G. T. Bartoo, J. D. Murray, E. C. Alvord Jr, *A mathematical model of glioma growth: the effect of chemotherapy on spatio-temporal growth,* Cell Proliferat. **28** (1995) 17–31.

[25] T. A. Ulrich, E. M. de Juan Pardo, S. Kumar, *The Mechanical Rigidity of the Extracellular Matrix Regulates the Structure, Motility, and Proliferation of Glioma Cells,* Cancer Res. **69** (2009) 4167–4174.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2014 3–7July, 2014.

Accurate and Efficient Electronic Structure Modeling of Organic Molecular Crystals

Jan Gerit Brandenburg[1] **and Stefan Grimme**[1]

[1] *Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie, Rheinische Friedrich-Wilhelms Universität Bonn, Beringstraße 4, 53115 Bonn, Germany*

emails: gerit.brandenburg@thch.uni-bonn.de, grimme@thch.uni-bonn.de

Abstract

The demand for high accuracy and efficiency in lattice energy minimization challenges modern theoretical methods. This is for instance crucial for organic crystal structure prediction. We review the applicability of *ab-initio* and semi-empirical approaches on various gas phase and solid state databases. It is demonstrated that London dispersion corrected Density Functional Theory (DFT-D) is very accurate with deviation from the references of only 0.5-1.5 kcal/mol (5-10%). While DFT-D can in principle distinguish between different polymorphs, the computational demand e.g. to screen a huge number of structures is too high for routine application. This task can be carried out by semi-empirical methods. A dispersion corrected Density Functional Tight-Binding (DFTB-D) Hamiltonian shows promising results. The mean absolute deviations are approximately 2-3 time larger than for DFT-D at a speedup of two orders of magnitude. The results show how the semi-empirical method can be used complementary to ab-initio computations for pre-screening of numerous structures or to compute thermodynamic properties of large systems.

Key words: Dispersion Correction, Non-Covalent Interaction, Organic Crystals, Tight-Binding, Semi-Empirical MO, Density Functional Theory

1 Introduction

In order to accurately model molecules both in the gas and solid phase, the correct treatment of inter- and intramolecular interactions is mandatory. Due to a variety of applications the theoretical progress is an active research field.[1,2,3,4,5,6,7,8] While the short-ranged intramolecular forces can be described by semi-local density functionals (and approximations thereof),

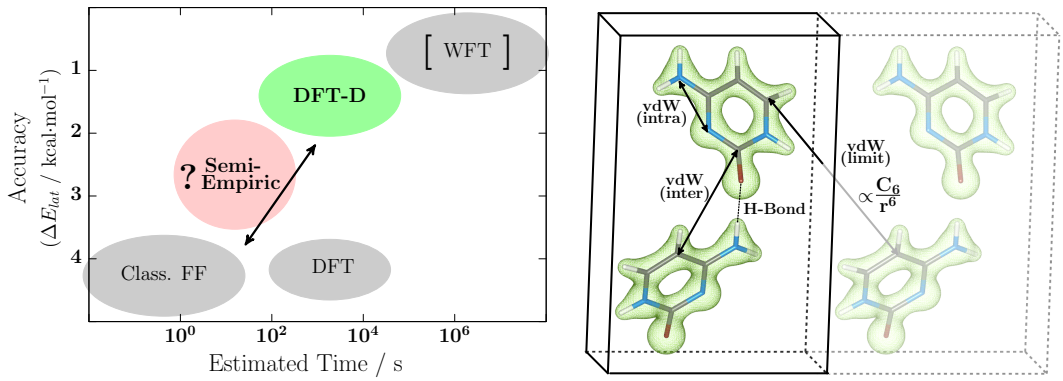


Figure 1: *Left:* A schematic view on the accuracy–computational cost ratio for different methods is given. The accuracy is exemplary given for the calculation of organic crystal lattice energies. Wavefunction theory methods (WFT) are expected to give the correct result in principle but can not be applied routinely. The gap between force fields and DFT-D is highlighted. *Right:* A typical molecular crystal with an electron density isosurface (calculated via DFT) is shown. The dominant intra- and intermolecular interactions are highlighted. *Reprint from Ref.*[4].

for gas phase dimers, supramolecular host-guest complexes, and organic crystals the long-range (non-covalent) forces are crucial. Most important of these intermolecular forces are the hydrogen bonding and van-der-Waals (vdW) interactions.[9] All local and non-local interactions are in principle described by high level quantum chemical methods. Although some efforts are made to apply localized versions of these methods, they are not applicable to very large complexes or to molecular crystals.[10] The mainly used alternatives are purely empirical potentials (force fields) and London dispersion corrected Density Functional Theory (DFT-D).[11] The non-local correlation can be incorporated by different means. G. Beran proposed a fragment-based hybrid many-body interaction model, which is capable to calculate lattice energies with chemical accuracy.[12,2] We recently demonstrated the predictive power of the semi-classical DFT-D3 scheme for molecular complexes and organic solids[13,14,15]. We calculated lattice energies of organic crystals with an accuracy of 1 kcal/mol.[3,16] Similarly accurate results are obtained with the Tkachenko-Scheffler (TS) Many-Body-Dispersion correction (MBD)[17,18] and E. Johnson's exchange dipole model (XDM)[19]. These methods have already been used in the field of crystal structure prediction.[20] Very recently, we combined a density functional tight-binding method with the D3 dispersion correction and evaluated the method on various benchmark sets.[4] In Figure 1, we show a typical molecular crystal and the (estimated) statistical accuracy of different methods. Wavefunction theory methods (WFT) are expected to give the correct result in principle but can not be applied routinely. On the other hand, empirical force fields have a too low accuracy and can only be used in an

 ISBN: 978-84-616-9216-3

on-the-fly parametrization scheme.[21] DFT-D is the method of choice if an accuracy of about 1 kcal/mol is needed. For pre-sceening techniques, semi-empirical models like DFTB-D3 can be conducted with accuracy of 2 to 3 kcal/mol.

Here, we want to put the possible multi-level approach into perspective. In section 2, we describe the utilized methods with focus on the London dispersion correction. We give a summary of the computational details in section 3. We compare and discuss the results of ab-initio DFT methods with different semi-empirical approaches on various benchmark sets (section 4). Finally, we conclude in section 5 with a short summary and outlook.

2 Method

At short distances, standard (semi local) density functionals can describe the effective electron interactions rather well. These interactions are closely related to change in the electron density and can therefore be modeled in a local expansion. However, non-local electron correlation cannot be described in this way. Therefore, the density functionals have to be corrected. The electron correlation between two fragments A and B at long distances r_{AB} can be connected to their dynamic polarizabilities at imaginary frequencies $\alpha(i\omega)$ (Casimir-Polder relation)[22,23]

$$E_{corr}^{AB}(r_{AB} \to \infty) = E_{disp}^{AB} = -\frac{3}{\pi}\int_0^\infty \alpha^A(i\omega)\alpha^B(i\omega)\,\mathrm{d}\omega \times \frac{1}{{r_{AB}}^6}\,. \tag{1}$$

The correct $1/r^6$ limit can not be described by semi-local density functionals, because the overlap of the electron density decays exponentially. This long range electron correlation, a.k.a. London dispersion interaction, can be added to the (semi-) local correlation captured by the density functional:

$$E_{total} = E_{dft} + E_{disp} \tag{2}$$

The standard DFT-D3 correction calculates the London dispersion energy in an atom-pairwise fashion

$$E_{disp} = -\frac{1}{2}\sum_{n=6,8}\sum_{i,j}^{N} s_n \frac{C_n^{ij}}{\|\mathbf{r}_{ij}\|^n + f(R_0^{ij})^n}\,, \tag{3}$$

where $C_{6/8}^{ij}$ are the leading order dipole–dipole and dipole–quadrupole dispersion coefficients and r_{ij} is the distance between the atom pairs i, j.[24] The s_6 scaling coefficient is set to unity to ensure the correct long-range behavior. The Becke-Johnson[25] rational damping function $f(R_0^{ij})$ is used to match the long- and medium-range dispersion contribution from D3 with the semi-local correlation captured by the density functional.[26] The C_6 dispersion

 ISBN: 978-84-616-9216-3

coefficients depend geometrically on the molecular environment and are pre-calculated by time-dependent DFT and utilizing the Casimir-Polder relation.

Complementary to the full ab-initio DFT calculations, we also utilize the Density Functional Tight-Binding method DFTB3. This method is based on a third-order expansion of the Kohn-Sham total energy with respect to charge density fluctuations. The arising matrix elements are modified by a self-consistent charge (SCC) redistribution. The modification corresponds to an on-site repulsion for short distances and to a Coulomb interaction at long distances with correct Coulomb limit. In the latest version an additional damping of the pair interactions involving hydrogen atoms is included. This significantly improves the description of hydrogen bonded systems and proton transfer.[27,28,29,30] We abbreviate this SCC-DFTB3 method as DFTB throughout the article. Similar to the DFT methods, the DFTB Hamiltonian has to be augmented with a London dispersion correction. Because the charge density of the DFTB method is (mainly due to its minimal basis) not very accurate, it is ideal to use a correction scheme which does not explicitly depend on the electronic structure.[4,31] The D3 correction solely uses the geometry information to calculate the dispersion energy. Because of its small numerical complexity, the D3 correction is ideally suited for a coupling with inherently fast electronic structure methods.

3 Computational Details

We calculate the PBE[32] (DFT) energy in large basis sets (def2-QZVP[33] and 1000 eV PAW[34]) using the TURBOMOLE 6.4[35] and the VASP 5.3[36,37] program suitye, respectively. The DFTB Hamiltonian with full third-order correction and self consistent charges (SCC) is computed via the `dftb+` standalone. We use the most recent Slater-Koster files provided by the group of M. Elstner. The hydrogen containing pair potentials are damped with an exponent of 4.2, which is the recommended value for proton transfers.[27,29,30] The PM6-DH2, PM7, and OM2 energies are calculated with the Mopac 2012 program[38] and the MNDO 7.0 program[39,40], respectively. The Brillouin zone is sampled with a Γ centered grid with at least 0.05 $Å^{-1}$ k-points, generated via the Monkhorst-Pack scheme.[41] The London dispersion correction D3 is used in the Becke-Johnson damping variant via the `dftd3` code.[24] The crystal geometries are optimized with fixed unit cell with the approximate normal coordinate rational function optimizer ANCOPT[42,43] until the atomic forces are below 10^{-4} au. For all other benchmarks the standard single-point energy approach was applied. In the X40 test set, systems including Br or I are excluded, and the Fe-containing complex in the S12L set is also disregarded due to missing Slater-Koster files.

 ISBN: 978-84-616-9216-3

4 Results

In order to validate the proposed PBE-D3 and DFTB-D3 method, we apply them to various standard databases and compare the results with the corresponding reference energies, non-dispersion corrected methods, and other semi-empirical methods. As prototypical density functional the widely used non-empiric PBE functional is used. Other generalized gradient approximated density functionals perform similar, while hybrid functionals perform slightly better.[13,16] Additionally to the DFTB model, we conduct the PM6, PM7, and OM2 methods as widely used semi-empirical Hartree-Fock approximation by neglect of diatomic differential overlap (NDDO).[44] We investigate the benchmark sets S22 (small gas phase dimers[45]), S66x8 (medium sized gas phase dimers at eight center-of-mass distances[46]), X40 (halogenated gas phase dimers[47]), L7 (large gas phase dimers and trimers[48,49]), S12L (large host-guest complexes[50]), and X23 (organic molecular crystals[19,17,18]). These data points are partially published elsewhere, see e.g.[3,4,14,48,13]. The results are summarized in Table 1.

Table 1: Mean absolute deviation (MAD), mean deviation (MD), and standard deviation (SD) of the dissociation and lattice energies for various benchmark sets are shown. Data are given for uncorrected as well as dispersion corrected (suffix D3) methods. All values are in kcal/mol and a positive MD denotes on average overbinding.

Method	MAD	MD	SD	MAD	MD	SD	MAD	MD	SD
	S22			**S66x8**			**X40**		
PBE	2.61	-2.58	3.74	1.52	-1.49	2.24	0.98	-0.92	1.89
PBE-D3	0.58	0.11	0.79	0.35	0.24	0.48	0.48	0.31	0.59
DFTB	3.50	-3.50	4.23	2.17	-2.17	2.54	2.12	-1.09	2.99
DFTB-D3	0.95	-0.80	1.56	0.79	-0.24	1.14	1.66	0.14	2.56
PM6	3.41	-3.41	4.22	2.00	-2.00	2.50	2.41	-2.41	3.70
PM6-DH2	0.39	-0.15	0.53	0.52	-0.26	0.80	1.63	-1.50	3.16
PM7	0.77	0.04	0.91	0.73	-0.13	0.96	1.69	-1.01	3.22
OM2-D3	0.93	-0.86	1.44	0.78	-0.44	1.19	—	—	—
	L7			**S12L**			**X23**		
PBE	15.59	-15.59	17.93	23.75	-23.75	28.50	11.70	-11.70	6.10
PBE-D3	1.58	0.26	1.63	2.01	1.21	2.53	1.07	0.43	1.34
DFTB	14.15	-14.15	15.95	19.79	-19.79	22.11	12.29	-12.29	13.58
DFTB-D3	1.74	1.31	2.28	5.90	4.60	7.99	2.48	-0.22	2.87
PM6	10.93	-10.93	12.84	14.36	-14.36	16.64	—	—	—
PM6-DH2	3.34	3.34	4.74	7.21	7.21	8.63	—	—	—
PM7	7.61	7.61	8.33	17.51	17.51	21.07	—	—	—
OM2-D3	2.36	-0.72	2.70	5.55	5.54	7.81	—	—	—

The failure of the non-dispersion corrected methods for these (non-covalently) bound systems is apparent. PBE, DFTB, and PM6 significantly underbind all molecular and periodic systems. This is not surprising as they are mainly bound by non-local correlation,

which cannot be captured by the utilized method. Augmenting the methods with the D3 dispersion correction (DH2 is a combined D2 dispersion correction and empirical hydrogen bonding correction and PM7 intrinsically has an empirical dispersion correction), reduces all MADs significantly. PBE-D3 has the lowest MAD with very accurate energies for all systems. The larger systems have slightly larger deviations, but this is due to the larger absolute interaction energy. The MAD of 1.1 kcal/mol for the molecular crystals X23 should be pointed out, it is below the estimated experimental error of 1.2 kcal/mol.[3] The dispersion corrected semi-empirical methods have MADs approximately a factor 2-5 larger. On the other hand, they are about two orders of magnitude faster. They also profit from the dispersion correction, but perform worse for the hydrogen bonded systems with the more complicated intermolecular electrostatic and induction. This is not surprising as this is the most crucial point for all semi-empirical approximations and can be seen at the larger MADs for the X40 set of halogenated dimers. Comparing the different semi-empirical methods, the dispersion corrected density functional tight-binding model DFTB-D3 seems to yield the most stable results. The MAD below 1 kcal/mol for the S66x8 set and of 2.5 kcal/mol for the X23 crystals is remarkable. The PM6-DH2 and OM2-D3 methods can also be recommended. However, its performance for the larger L7 and S12L systems is slightly worse than for the well balanced DFTB-D3. While PM7 has small MADs close to its fit sets S22 and S66x8, the results for the larger systems are far worse and it should be used with care.

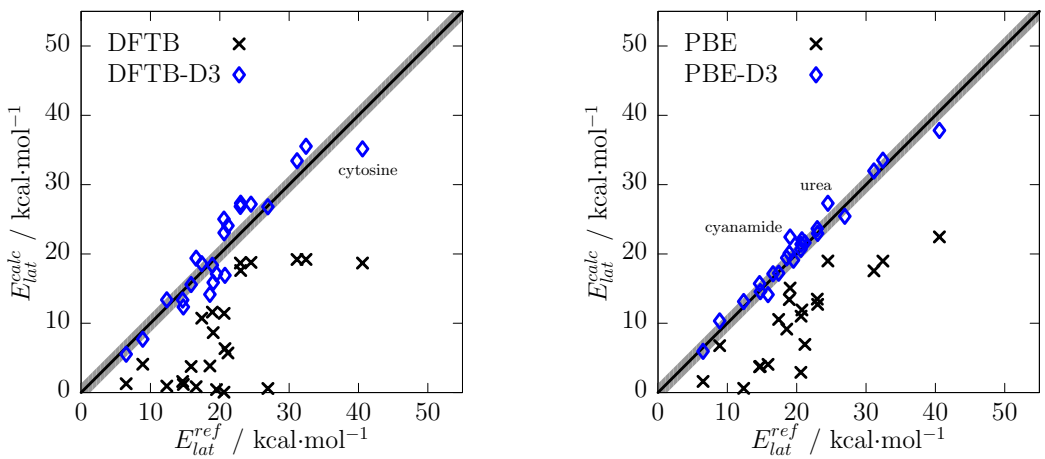


Figure 2: Correlation between the calculated DFTB, DFTB-D3, PBE, and PBE-D3 lattice energies with the experimental reference values. The gray shading denotes the experimental uncertainty of approximately 1.2 kcal/mol. *Reprint from Ref.*[4].

In Figure 2, we show the correlation between the calculated and reference lattice energies on the X23 set of organic crystals. Here, we focus on the dispersion corrected and

 ISBN: 978-84-616-9216-3

uncorrected PBE and DFTB methods. The uncorrected models neglect a main contribution to the lattice energy as shown by the large deviations from the references. This is efficiently corrected with the D3 dispersion correction. PBE-D3 and DFTB-D3 seem to be well suited to describe both gas and solid phase systems. The linear correlation coefficient is 0.98 and 0.94, respectively.

5 Conclusion

We presented and evaluated different electronic structure methods with focus on dispersion corrected DFT and semi-empirical methods. The MAD of the PBE-D3 scheme for binding energies is below 1 kcal/mol for the smaller benchmark systems S22, S66x8, and X40, approximately 2 kcal/mol for the larger L7 and S12L systems, and below the experimental error of 1.2 kcal/mol for the lattice energies of the X23 crystals. While DFT-D has the best accuracy, it is computationally too demanding for the screening of numerous structures. Semi-empirical methods can fill this gap, as shown by the small MADs of DFTB-D3, i.e. 0.9, 0.8, 1.7, 1.7, 5.9, and 2.5 kcal/mol for the S22, S66x8, X40, L7, S12L, and X23 test sets. PM6-DH2 can also be recommended, however, the errors for the larger systems are significantly larger compared to DFTB-D3.

Interestingly, the relative errors for both the molecular dimers and the molecular crystals are very similar. The constantly good performance of the PBE-D3 and DFTB-D3 methods demonstrates their robustness. A combined usage of both methods seems to be ideal for the challenging task of organic crystal structure prediction. Another possible application is the calculation of sublimation energies, combining the electronic energy from PBE-D3 with the free energy correction from DFTB-D3. This multi-level approach is already routinely done for gas phase systems, but could be extended without modification to the solid state.

References

[1] A. J. Cruz-Cabeza and J. Bernstein. Conformational polymorphism. *Chemical Reviews*, 114:2170–2191, 2014.

[2] K. Nanda and G. Beran. Prediction of organic molecular crystal geometries from MP2-level fragment quantum mechanical/molecular mechanical calculations. *J. Chem. Phys.*, 138:174106, 2012.

[3] J. G. Brandenburg and S. Grimme. Dispersion Corrected Hartree-Fock and Density Functional Theory for Organic Crystal Structure Prediction. *Top Curr Chem*, 2013. DOI: 10.1007/128_2013_488.

[4] J. G. Brandenburg and S. Grimme. Accurate Modeling of Organic Molecular Crystals by Dispersion-corrected Density Functional Tight Binding (DFTB). *J. Phys. Chem. Lett.*, 2014. DOI: 10.1021/jz500755u.

[5] S. M. Woodley and R. Catlow. Crystal structure prediction from first principles. *Nature Materials*, 7:937–964, 2008.

[6] G. J. O. Beran, S. Wen, K. Nand, Y. Huang, and Y. Heit. Accurate and Robust Molecular Crystal Modeling Using Fragment-Based Electronic Structure Methods. *Top. Curr. Chem.*, 2013. DOI: 10.1007/128_2013_502.

[7] C. C. Pantelides, C. S. Adjiman, and A. V. Kazantsev. General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules. *Top. Curr. Chem.*, 2014. DOI: 10.1007/128_2013_497.

[8] A. R. Oganov. *Modern Methods of Crystal Structure Prediction*. Wiley-VCH, Berlin, 2010.

[9] A. J. Stone. *The Theory of Intermolecular Forces*. Oxford University Press, Oxford, 1997.

[10] C. Riplinger and F. Neese. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.*, 138:034106, 2013.

[11] S. Grimme. Density functional theory with london dispersion corrections. *WIREs Comput. Mol. Sci.*, 1:211–228, 2011.

[12] G. J. O. Beran and K. Nanda. Predicting organic crystal lattice energies with chemical accuracy. *J. Phys. Chem. Lett.*, 1:3480–3487, 2010.

[13] L. Goerigk and S. Grimme. Efficient and accurate double-hybrid-meta-gga density functionals-evaluation with the extended gmtkn30 database for general main group thermochemistry, kinetics, and noncovalent interactions. *J. Chem. Theory Comput.*, 7:291–309, 2011.

[14] T. Risthaus and S. Grimme. Benchmarking of london dispersion-accounting density functional theory methods on very large molecular complexes. *J. Chem. Theory Comput.*, 9:1580–1591, 2013.

[15] J. G. Brandenburg, A. Hansen, S. Grimme, H. Eckert, and G. Erker. Crystal packing induced carbon-carbon double-triple bond isomerization in a zirconocene complex. *J. Am. Chem. Soc.*, 2014. DOI: submitted.

 ISBN: 978-84-616-9216-3

[16] J. Moellmann and S. Grimme. A DFT-D3 Study of Some Molecular Crystals. *J. Phys. Chem. A*, 2014. DOI: 10.1021/jp501237c.

[17] A. M. Reilly and A. Tkatchenko. Seamless and Accurate Modeling of Organic Molecular Materials. *J. Phys. Chem. Lett.*, 4:1028, 2013.

[18] A. M. Reilly and A. Tkatchenko. Understanding the role of vibrations, exact exchange, and many-body van der waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.*, 139:024705, 2013.

[19] A. Otero de-la Roza and Erin R. Johnson. A benchmark for non-covalent interactions in solids. *J. Chem. Phys.*, 137:054103, 2012.

[20] David A. Bardwell, Claire S. Adjiman, Yelena A. Arnautova, Ekaterina Bartashevich, Stephan X. M. Boerrigter, Doris E. Braun, Aurora J. Cruz-Cabeza, Graeme M. Day, Raffaele G. Della Valle, Gautam R. Desiraju, Bouke P. van Eijck, Julio C. Facelli, Marta B. Ferraro, Damian Grillo, Matthew Habgood, Detlef W. M. Hofmann, Fridolin Hofmann, K. V. Jovan Jose, Panagiotis G. Karamertzanis, Andrei V. Kazantsev, John Kendrick, Liudmila N. Kuleshova, Frank J. J. Leusen, Andrey V. Maleev, Alston J. Misquitta, Sharmarke Mohamed, Richard J. Needs, Marcus A. Neumann, Denis Nikylov, Anita M. Orendt, Rumpa Pal, Constantinos C. Pantelides, Chris J. Pickard, Louise S. Price, Sarah L. Price, Harold A. Scheraga, Jacco van de Streek, Tejender S. Thakur, Siddharth Tiwari, Elisabetta Venuti, and Ilia K. Zhitkov. Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. *Acta Crystallographica Section B*, 67:535–551, 2011.

[21] M. A. Neumann, F. J. J. Leusen, and J. Kendrick. A major advance in crystal structure prediction. *Angew. Chem. Int. Ed.*, 47:2427–2430, 2008.

[22] H. B. G. Casimir and D. Polder. *Phys. Rev.*, 4:360–372, 1948.

[23] S. J. A. van Gisbergen, J. G. Snijders, and E. J. Baerends. A density functional theory study of frequency-dependent polarizabilities and van der waals dispersion coefficients for polyatomic molecules. *J. Chem. Phys.*, 103:9347–9354, 1995.

[24] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *J. Chem. Phys.*, 132:154104, 2010.

[25] E. R. Johnson and A. D. Becke. *J. Chem. Phys.*, 124:174104, 2006.

[26] S. Grimme, S. Ehrlich, and L. Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32:1456–1465, 2011.

[27] M Elstner, D Porezag, G Jungnickel, J Elsner, M Haugk, T Frauenheim, S Suhai, and G Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B*, 58(11):7260–7268, 1998.

[28] B. Aradi, B. Hourahine, and Th. Frauenheim. Dftb+, a sparse matrix-based implementation of the dftb method. *J. Phys. Chem. A*, 111(26):5678–5684, 2007.

[29] M. Elstner. SCC-DFTB: What Is the Proper Degree of Self-Consistency? *J. Phys. Chem. A*, 111:5614–5621, 2007.

[30] M. Elstner. The SCC-DFTB method and its application to biological systems. *Theor. Chem. Acc.*, 116:316–325, 2006.

[31] M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, and E. Kaxiras. *J. Chem. Phys.*, 114:5149–5155, 2001.

[32] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, 1996. erratum *Phys. Rev. Lett.* **78**, 1396 (1997).

[33] F. Weigend, F. Furche, and R. Ahlrichs. *J. Chem. Phys.*, 119:12753–12762, 2003.

[34] P. E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50:17953, 1994.

[35] TURBOMOLE 6.4: R. Ahlrichs et al., Universität Karlsruhe 2009. See http://www.turbomole.com.

[36] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *J. Comp. Mat. Sci.*, 6:15, 1996.

[37] G. Kresse and J Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169, 1996.

[38] MOPAC2012, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (2012).

[39] W. Thiel and A. A. Voityuk. *J. Phys. Chem.*, 100:616, 1996.

[40] MNDO2005 Version 7.0, W. Thiel, MPI für Kohlenforschung, Mülheim, Germany.

[41] H. J. Monkhorst and J. D. Pack. Special points for Brillouin-zone integrations. *Phys. Rev. B*, 13:5188–5192, 1976.

[42] S. Grimme, ANCOPT: Approximate Normal Coordinate Rational Function Optimization, University of Bonn 2014.

 ISBN: 978-84-616-9216-3

[43] F. Eckert, P. Pulay, and H.-J. Werner. Ab initio geometry optimization for large molecules. *J. Comput. Chem.*, 18:1473–1483, 1997.

[44] J. J. P. Stewart. Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *J. Mol. Mod.*, 13:1173, 2007.

[45] P. Jurečka, J. Šponer, J. Cerny, and P. Hobza. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.*, 8:1985–1993, 2006.

[46] J. Řezáč, K. E. Riley, and P. Hobza. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.*, 7:2427, 2011.

[47] J. Řezáč, K. E. Riley, and P. Hobza. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *J. Chem. Theory Comput.*, 8:4285–4292, 2012.

[48] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay, and P. Hobza. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *J. Chem. Theory Comput.*, 9:3364–3374, 2013.

[49] A. Hansen and S. Grimme. Revised l7 references on the dlpno-ccsd(t)/Δcbs/cp level. To be published.

[50] S. Grimme. Supramolecular binding thermodynamics by dispersion corrected density functional theory. *Chem. Eur. J.*, 18:9955–9964, 2012.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

On methodological aspects of traffic theory

Buslaev A.P.[1] **and Gorodnichev M.G.**[2]

[1] *Department of Mathematics, Moscow Automobile and Road University*

[2] *Department of Mathematical Cybernetics and Information Technologies, Moscow Technical University of Communication and Informatics*

emails: Apal2006@yandex.ru, gorodnichev@hotmail.com

Abstract

This article is about methodological aspects of traffic theory. In this article we considered new and old model of traffic flow and their advantages and disadvantages.

Key words: traffic flow, methodological, congested traffic flow
MSC 2000: AMS codes (optional)

1 Introduction

In the 20s of the 20th century in industrialized countries began the process of motorization. Vehicles are becoming more affordable, comfortable and increasing traffic flow should be controlled. Therefore, the scientists faced the task of investigating the behavior of traffic flow and the development of methods for the prediction and management.

Created in 20th century traffic flow theory divided into microscopic and macroscopic.

(1) Microscopic approach considers the behavior of individual pairs of cars, moving one after another without overtaking. This approach is called microscopic. This approach developed Grindshields [1], Pipes [3], Newell [5] and others.

(2) The macroscopic approach considers the traffic flow as a whole. The first macroscopic model was obtained Grindshields [2], later-model Lighthill - Whitham [7], etc.

Over the past 80 years have created many variants of models (1) - (2). Considered one of the main model is car-following model. However, in some cases in the car-following model jumps are observed basic characteristics of vehicular traffic called "breakdown phenomenon".In their studies these phenomena encountered Pipes [4], Newell [6], Herman [10], Gazis [8], Rothery [9], etc.

2 About Kerner's three-phase traffic theory

In 00s years of the 21th century was devised Kerner's three-phase traffic theory [11], with which the author claims [12, page.v], we can explain and predict the spatiotemporal empirical features of traffic breakdown.

Let's review three main concepts of the theory.

Free traffic flow, [12, page 257] is usually observed, when the vehicle density in traffic is small enough The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function. The increase in the flow rate with the density increase in free flow has a limit. At the associated limit (maximum) point of free flow, the flow rate and density reach their maximum values while the average speed has a minimum value that is still possible in free flow.*, [12, page 255]* - Congested traffic can be defined as a state of traffic in which the average speed is lower than the minimum average speed that is still possible in free flow.

Front of traffic pattern, [12, page 257] is either a moving or motionless region within which one or several of the traffic variables change abruptly in space (and in time, when the front is a moving one). There are downstream front and upstream front of the traffic pattern. *The downstream pattern front, [12, page 257]* separates the pattern from other traffic patterns downstream. *The up stream pattern front, [12, page 257]* separates the pattern from other traffic patterns upstream.

Wide moving jam, [12, page 262] In three-phase theory, the following definition [J] of the wide moving jam traffic phase in congested traffic is made. A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front? even when the jam propagates through any other traffic states or bottelnecks. This is the characteristic feature [J] of wide moving jam phase.

Narrow moving jam, [12, page 259] is a moving jam, which consists of jam fronts only. Narrow moving jams are associated with the synchronized flow phase.

Synchronized flow phase, [12, page 261] in three-phase theory, the following definition [S] of the synchronized flow phase in congested traffic is made. In contrast to the wide moving jam phase, the downstream front of the synchronized flow phase does not maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often at a bottleneck. In other words, synchronized flow does not exhibit the characteristic jam feature [J].

The term *synchronized flow* reflects the following features of this phase of the traffic:

1) *The continuous flow* without significant *delays* which often occur within a wide moving cluster.

2) There is a tendency to synchronize speeds in the stream. Furthermore, there is a tendency to synchronization of vehicles on each of the lanes of the road (the formation of groups of vehicles) in a synchronized stream.

 ISBN: 978-84-616-9216-3

Synchronized flow unlike widely moving jam does not retain its average speed at the downstream front.

Bottleneck, [12, page 255] The breakdown phenomenon leading to the onset of traffic congestion occurs mostly at a highway bottleneck. On average the speed is lower and density is greater within this disturbance than these traffic variables are in free flow outside of the disturbance.

Three-phase traffic theory makes the assumption that in addition to free traffic flow phase *has two phase of congested traffic: synchronized flow and wide moving jam.* Thus, there are three phases in a three-phase traffic theory: *free flow (F); synchronized flow (S); wide moving jam (J).*

Let's review the methodology of the theory (p. 103). Spatiotemporal measurements of the traffic flow features are taken on different highways in different countries for long period of time. In that data clear features of intense traffic flow structure are found as well as generic features. On the next stage microscopic and macroscopic criteria are defined for different phases of the three phase theory, common for different structures of traffic , common structures are defined for only intense traffic phase, certain choice of parameters is valid for each phase. Examples of such class of structures for intense traffic with common qualitative spatiotemporal features are structures , appearing in bottleneck.

Therefore, empirical data lie at the base of Kerner theory without mathematical formulation and models [11, . 87] about density, intensity, averaged velocity which were collected with detectors. *Formulations given in [11]*, don't have algorithm on how to identify reviewed objects, therefore the conclusions can't have serious consequences. This probably prompted the author to expand his ideas in [12].

3 Flow as the composite of collective and individual

Description of traffic flow behavior via measuring density, intensity and averaged velocity is clearly not sufficient for such a complicated process, what factually Kerner [11] tries to state, using the modern data monitoring. Classic models simplifying traffic flow as dependency between velocity and intensity, well proven in hydrodynamic as example, are not valid for traffic, where dozens of particles interact (not millions) on each kilometer of lane, who also have their own pattern of behavior. For sure the base concept in traffic is the concept of dynamic clearance limit which represents untouched space for the vehicle on the road. The simplest model of dynamic dimension is quadratic dependency on velocity [1] with coefficients dependent on several indicators of low formalized complex human nature - this is how the society obeys to God and not mechanics of Newton. With the first approach these coefficients can be considered as constants, velocity is the recommended value for the concrete conditions of the flow, dynamic clearance limit is the size of the region equal to one lane, that is necessary for vehicles' safe movement.

Therefore, dynamic dimension limit helps to divide the road space on cells *in order to define the flow connection correctly as the percent of busy cells on given speed mode [13].* Current state of the driver can be taken *as probability to move to neighboring cell within period of time.* This approach unites determinate and episodic models that were widely presented in the history of traffic.

4 Cluster as a steady state of a connected chain of follow the leader

Model of one lane connecting movement takes us to the system of lineal differential equations in ideal alternative.

$$x_{n+1}(t) - x_n(t) = f_n(\dot{x}_n(t)) \tag{1}$$

where f - the safety function, the dynamic dimension. Human factor in the right part is considered in different ways of that approach: *(a) delay of argument - time for the driver to validate the information; (b) precision with which the distance to the car moving ahead was evaluated [8, 10, etc.].*. Modern tendencies in car systems take it to the state where "occasional variations of human behavior" become of low importance while the cars are getting "smarter".

With very generic assumptions [14] it is correct to state stability of equal chain spread if the leader follows with equal velocity. As a result in canalized traffic the sequence of equally set vehicles, moving with equal velocity is found - it cluster, where density statement is correct, and velocity is one value function from density in case of monotonous functions of dynamic clearance limit [17].

Praising hydrodynamic approach in traffic, military terminology can be widely used - as Kerner does, used along with *downstream front, upstream, etc.* Interaction of waves can be defined via different methods, one of which is local - analog to conditions of Gugono-Renkin and takes in reviewed case to the system ODE with changing architecture [17], [16].

5 Cluster models of multiband movement

Red Color - holy color for inhabitants former USSR at of that time all atheism. A jumble of notions produced in two *red books* [11], [12] will try to systematize. Thus, the cluster is sustainable formation of the vehicles flow is evenly spread of particles having a constant speed at a fixed density [15]. Cluster remains in unchanging configuration unless enters interaction with other clusters or if disrupted geometry of the road. Interaction of the clusters described by a system of ordinary differential equations - it is one of the variants of behavior of a totally connected movement. In this "front line" depending on the intensity

 ISBN: 978-84-616-9216-3

of the flow on the leading front of outsider and the trailing front of the leading can shuffle in any direction.

$$\dot{x} = \frac{v_2\rho_2 - v_1\rho_1}{\rho_2 - \rho_1} \tag{2}$$

Model (2) - analog of the Hugoniot-Rankine, $v_2 = v_2(\rho_2)$, $v_1 = v_1(\rho_1)$, [17].

Multiband movement further comprises overtaking procedure and its variations: *overtaking, temporizing* - quick follow close behind the with slow speed slow, *humility* - to slow the rapid integration according to (2), *follow* - slow to fast integration with a totally connected movement, which is an alternative to lag [16].

The general problem of cluster modeling reduces to the following. There is some initial configuration of the cluster length l_i, density y_i, "fronts" - borders $[a_i, b_i]$, $i = 1, ..., n$ on several bands $X^{(i)}$, $j = 1, ..., m$. During the movement more rapid clusters overtake slow and starts the process of interaction based on a set of the above procedures. The purpose of the study is to evaluate the configurations of the cluster fields and numerical flow characteristics. In particular, the presence of bottlenecks.

Cluster model of motion is a special case of general deterministic-stochastic approach, assuming a totally connected behavior [17].

6 Conclusion

In the works of Daganzo [18] is the initial attempt to break a vicious circle: at first introduced weakly defined macrocharacteristics - density and velocity, and then for them issued fanciful equations in partial derivatives, which in turn is necessary to investigate numerically, ie using finite difference schemes. Schreckenberg and Nagel [19] continued the work renunciation of of attempts describing the behavior traffic through the classical and generalized solutions PDEs, forming a reasonable statement immediately at a discrete level - it is announced to us the theory of cellular automata [19] finally, *a modern agent-based modeling* , armed with supercomputers, claims to describe the behavior of hundreds of thousands of vehicles in parallel. This is the other extreme - the polar opposite to the fans to evaluate performance in uncertain fashion equations of mathematical physics. If not limited by a finite set of parameters and transparent rules of behavior models, it is not possible to estimate the number of different variants of behavior of socio-technical system, for example, hundreds of thousands of cars on the road network of the metropolis, agent-based modeling so well in only two cases: *long to simulate in a tiny place, or the entire road network, but a very short time.* As time has shown, *military terminology physicists - wave fronts*, etc. also can not do. Optimum inspires rapid development of information and communication technologies that reduce the number of degrees of freedom considered STS. Therefore, *restricting freedom of choice and the chaos of thoughts* drivers on the one hand and *surgically gently introducing new approaches to modeling* with measurable (and not smearing) parameters,

 ISBN: 978-84-616-9216-3

you can achieve success in managing and saturated traffic flow on the street and complex road networks.

7 Mathematical problems of of modern approaches to modeling traffic

We formulate basic directions of mathematical research

(7.1) The standardization of procedures for creating multilevel model of of the road network

(7.2) Circular planar graphs: regular and quasiregular network

(7.3) Monotone wander around the ring networks: rules, numerical characteristics and dynamics of congested traffic

(7.4) Cluster model for regular periodic networks. Compressible and rigid clusters

(7.5) Management at networks. Optimization of traffic on the specified criteria

All of the above corrections are sources strictly set of mathematical problems related to the different sections: the theory of dynamical systems, Markov processes, the theory of differential equations, the theory of functions [15], [13], [14].

Furthermore, it seems obvious to use these approaches to the fields of natural sciences and engineering, non-road. Large intersection viewed from cellular automata theory (Komvey etc.), some models of biological areas. Computer implementation and research of such problems leads to problems of parallel computing, and other methods of increasing the efficiency of computer algorithms.

References

[1] B. D. Greenshields, *The Photographic Method of Studing Traffic Behavior Highway*, RES. Board Proc. 1933,v.13

[2] B. D. Greenshields, *A Study of Traffic Capacity*, Proceedings of the Highway Research Board, 14, 468, 1935.

[3] L. A. Pipes, *A Proposed Dynamic Analogy of Traffic*, ITTE Report, Institute of Transportation and Traffic Engineering, University of California, Berkeley, 1951.

[4] L. A. Pipes, *An Operational Analysis of Traffic Dynamics*, Journal of Applied Physics **24** (1953) 271-281.

[5] G. F. Newell, *Nonlinear Effects in the Dynamics of Car Following*, Operations Research **9(2)** (1961) 209-229.

 ISBN: 978-84-616-9216-3

[6] G. F. Newell, *Theories of Instability in Dense Highway Traffic*, Operations Research Society of Japan **5(1)** (1962) 9-54.

[7] M. J. Lighthill, G. B. Whitham , *On kinematic waves: Theory of traffic flow on long crowded roads*, Proc. R. Soc. London, Ser. A., 1955.

[8] D. C.Gazis, R. Herman, R. B. Potts, *Car Following Theory of Steady State Traffic Flow*, Operations Research **7(4)** (1959) 499-505.

[9] D. C.Gazis, R. Herman, R. W. Rothery , *Non-Linear Follow the Leader Models of Traffic Flow*, Operations Research **9** (1961) 545-567.

[10] R. Herman, R. W. Rothery , *Microscopic and Macroscopic Aspects of Single Lane Traffic Flow*, Operations Research, Japan (1962) pp. 74.

[11] B. S. Kerner, *The physics of traffic*, Springer, 2004, 683 p.

[12] B. S. Kerner, *Introduction to Modern Traffic Flow Theory and Control*, Berlin: Springer, 2009.

[13] A. P. Buslaev, A. V. Novikov, V. M. Prihodko, A. G. Tatashev, M. V. Yashina, *Stochastic simulation and optimization approaches to road traffic*, "world", 2003. - 368 .

[14] A. P. Buslaev, A. V. Gasnikov, M. V.Yashina , *Mathematical Problems of Traffic Flow Theory*, Proceedings of the 2010 International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE - 2010, Almeria (Andalucia), Spain (2010) 307-313.

[15] V. V. Kozlov, A. P. Buslaev, A. G. Tatashev, *Monotonic random walks and clusters flows on networks. Models and traffic applications*, Lambert Academic Publishing, 2013, 300 p.

[16] P. M. Strusinskiy, *Cluster model of traffic on multi-lane networks*, Vestnik MADI 2014

[17] A. P. Buslaev, P. M. Strusinskiy, *Computer simulation analysis of cluster model of totally-connected flows on the chain mail*, New Results in Dependability and Computer Systems, Proceedings of the 8th International Conference Dependability and Complex Systems DepCoS-RELCOMEX'13, Springer, Brunow, Poland 2013, 63-74.

[18] C. F. Daganzo, *Problem Sets: Fundamentals of Transportation and Traffic Operations*, Institute of Transportation Studies, University of California at Berkley, 1998

[19] K. Nagel, M. Schreckenberg, *A cellular automation model for freeway traffic*, Phys. I France, 1992, v.2

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

A computational study on the quickest path problem with energy constraints

Herminia I. Calvete[1], Lourdes del-Pozo[1] and José A. Iranzo[1]

[1] *Dpto. de Métodos Estadísticos, Universidad de Zaragoza*

emails: herminia@unizar.es, lpozo@unizar.es, joseani@unizar.es

Abstract

In this paper an extension of the quickest path problem is addressed in which an additional parameter is added representing energy at the nodes. This energy is needed to transmit the items through the quickest path. The aim of this constrained quickest path problem is to obtain a quickest path whose nodes are able to support the transmission of a message of known size. After formulating the problem, the computational performance of an algorithm developed to solve this constrained quickest path problem is assessed.

Key words: quickest path, energy constraint, shortest path
MSC 2000: 90B10, 90B18

1 Introduction

The quickest path problem (QPP) consists of finding a path in a directed network to transmit a given amount of items from an origin node to a destination node with minimal transmission time, when the transmission time depends on both the traversal times of the arcs, or lead time, and the rates of flow along arcs, or capacity. The QPP can be viewed as a variant of the well-known shortest path problem (SPP), but it is worth pointing out that the quickest path depends on both the characteristics of the network and the amount of items to be transmitted. The QPP was first proposed by Moore [10] to model flows of convoy-type traffic. Then, it was proposed by Chen and Chin [5] in the context of modeling transmission problems in communication networks. Later, Martins and Santos [9] approached the QPP as a special minsum-maxmin bicriteria path problem. Several polynomial time solution algorithms have been proposed in the literature based on transforming the original problem in solving either a shortest path problem in an enlarged network or a

sequence of shortest path problems [3, 5, 9, 11, 13, 15]. Furthermore, several variants and extensions of the QPP have been considered. Pascoal et al. [12] provide a survey on the subject. Calvete and del-Pozo [2] dealt with the QPP when items are transmitted in batches of variable size. Hamacher and Tijandra [7] propose the model in a special evacuation problem where evacuees may use only a single path or tunnel from their initial position. Clímaco et al. [6] develop an algorithm to compute K quickest paths in the context of internet packet routing. Calvete, del-Pozo and Iranzo [3] propose algorithms to compute the quickest path whose reliability is not lower than a given threshold. To deal with uncertain or imprecise coefficients, Calvete [1] proposes to introduce interval coefficients and take into account the decision maker's preference, whereas Ruzika and Thiemann [14] propose a discrete scenario technique.

In this paper we address the quickest path problem with energy constraints (EQPP) introduced by Calvete, del-Pozo and Iranzo [4]. In this variant of the QPP, an additional parameter is associated to nodes, which can be referred to as the battery energy available to transmit items. The aim of the EQPP is to obtain a quickest path whose nodes are able to support the transmission of a message of size σ. In [4] a polynomial time algorithm based on solving shortest path problems in properly defined networks is proposed, whose computational performance is assessed in this paper. Sections 2 and 3 formally set out the quickest path problem and the quickest path problem with energy constraints, respectively. Section 4 goes on to assess the computational performance of the algorithm.

2 Mathematical formulation of the quickest path problem

Let $\mathcal{G} = [\mathcal{N}, \mathcal{A}]$ be a directed network without multiple arcs and self loops, where $\mathcal{N}$ denotes the set of nodes and $\mathcal{A}$ the set of directed arcs. Let n be the number of nodes and m the number of arcs. Let s and t be two distinguished nodes in the network called, respectively, origin and destination and σ the amount of items to be sent from node s to node t. Each arc $a = (u, v) \in \mathcal{A}$ has two associated parameters: a capacity $c(u, v) \geq 0$, which represents the maximum number of items that can flow from node u to node v through arc (u, v) per unit of time, and a lead time $l(u, v) \geq 0$, which represents the time required for the items to traverse the arc (u, v). We assume that there are r different capacities $c_1 < c_2 < \ldots < c_r$.

A simple path or loopless path P from node s to node t is a sequence of nodes and arcs $P = (s = u_1, u_2, \ldots, u_k = t)$ such that $u_i \in \mathcal{N}$, $i = 1, \ldots, k$, $u_i \neq u_j$ if $i \neq j$, and $(u_i, u_{i+1}) \in \mathcal{A}$, $i = 1, \ldots, k-1$. In the paper, we use the term path in place of simple or loopless path for short as well as the term $s - t$ path in place of a path from s to t. The lead time along path P is

$$l(P) = \sum_{(u,v) \in P} l(u, v)$$

 ISBN: 978-84-616-9216-3

The capacity of the path is

$$c(P) = \min_{(u,v)\in P} c(u,v)$$

The transmission time required to send σ items from s to t along the path P is

$$T_P(\sigma) = l(P) + \left\lceil \frac{\sigma}{c(P)} \right\rceil, \tag{1}$$

where $\lceil\cdot\rceil$ denotes the ceiling function, i.e., $\lceil x\rceil$ is the smallest integer not less than x.

Let $\mathcal{P}$ be the set of $s-t$ paths in network $\mathcal{G}$. We assume that $\mathcal{P} \neq \emptyset$. The QPP can be formulated as finding an $s-t$ path so that

$$\begin{array}{ll} \min\limits_{P} & T_P(\sigma) \\ \text{s.t.} & P \in \mathcal{P} \end{array}$$

3 Mathematical formulation of the quickest path problem with energy constraints

Now, we assume that nodes represent transmitters/receivers and each node $u \in \mathcal{N}$ has an associated parameter b_u which represents a limited energy battery. When $\pi \in (0, c(u,v)]$ items are transmitted from node u to node v, the available energy at node u decreases by a constant amount which depends on both the distance between nodes u and v and the capacity $c(u,v)$, but which is independent of π. For notational convenience, let us assume that the distance between nodes u and v is proportional to the lead time $l(u,v)$ and the proportionality constant is 1. Therefore, the transmission energy required at node u to transmit $\pi \in (0, c(u,v)]$ items through the arc (u,v) is:

$$\omega(u,v) = \alpha_{c(u,v)} l^{\beta}(u,v)$$

where $\alpha_{c(u,v)}$ are constants verifying $0 < \alpha_{c_1} < \alpha_{c_2} < \ldots < \alpha_{c_r}$ and β is a parameter which depends on the network.

Let $P = (s = u_1, u_2, \ldots, u_k = t) \in \mathcal{P}$. Let $b_u(\sigma, P)$ denote the residual energy at node u after transmitting σ items through the path P. Hence,

$$b_u(\sigma, P) = \begin{cases} b_u - \omega(u_i, u_{i+1}) \left\lceil \frac{\sigma}{c(P)} \right\rceil & \text{if } u = u_i, i = 1, \ldots, k-1 \\ b_u & \text{otherwise} \end{cases}$$

From now on we assume without loss of generality that

$$b_u - \omega(u,v) \left\lceil \frac{\sigma}{c(u,v)} \right\rceil \geqslant 0, \forall (u,v) \in \mathcal{A} \tag{2}$$

 ISBN: 978-84-616-9216-3

If an arc does not meet the condition, this arc can be removed from the network since it will not be used in the transmission.

A path $P \in \mathcal{P}$ is said to be feasible with respect to the transmission of a message of size σ if $b_u(\sigma, P) \geqslant 0$, $\forall u \in P$. That is to say, the feasibility of a path is measured through the availability of its nodes to transmit the whole message. The EQPP can be formulated as:

$$\begin{array}{ll} \min\limits_{P} & T_P(\sigma) \\ \text{s.t.} & b_u(\sigma, P) \geqslant 0, u \in \mathcal{N} \\ & P \in \mathcal{P} \end{array}$$

In order to solve the problem, Calvete, del-Pozo and Iranzo [4] propose to use the algorithm EQPA based on successively solving shortest path problems in properly defined networks $\mathcal{G}_j$, $j = 1, \ldots, r$. Essentially, the network $\mathcal{G}_j$ maintains the arcs of $\mathcal{G}$ with capacity greater or equal c_j which would be able to transmit the message of size σ.

4 Computational performance of the algorithm EQPA

In order to analyze the performance of the algorithm, we have generated a set of test problems using a modified version of NETGEN [3, 8]. The skeleton provided by NETGEN is used. Lead time and capacity coefficients are generated from uniform distributions in the range $[1, 50]$. There are four main problem groups defined by the number of nodes $n = 1000$ and $n = 5000$ and the number of arcs $m = 10n$ and $m = 20n$. For each problem group, we generate problems having $r = 10$ and $r = 20$ distinct capacities. For this purpose, first the required number of capacities are generated from the corresponding uniform distribution. Then, we assign to each arc one of the capacities generated with a uniform probability.

For each arc $(u, v) \in \mathcal{A}$, the coefficient $\alpha_{c(u,v)}$ is assumed to be proportional to its capacity, with proportional constant $\alpha = 0.001$. For each node $u \in \mathcal{N}$, b_u is obtained multiplying $b = \max\limits_{(u,v)\in\mathcal{A}} \alpha c(u, v) l^{\beta}(u, v) \left\lceil \frac{5000}{c(u,v)} \right\rceil$ by a random number in the interval $[.75, 1.25]$.The parameter β takes the value 2. For assessing the effect of the number of items which are sent, a quickest path for $\sigma = 1000$, 5000, 20000 and 50000 is computed in each problem. For these values of σ, approximately 100%, 96%, 70% and 44% of the total number of arcs m in the original network verify (2). The numerical experiments have been performed on a PC Intel Pentium D CPU at 3.0 GHz having 3.2 GB of RAM under Ubuntu Linux 10.04. The code has been written in C++, GCC 4.4.3.

Tables 1 to 4 summarize the results of the experiment for each problem group. The upper part of each Table displays the results for $r = 10$. The results for $r = 20$ are shown in the lower part. The first column shows the network considered. Depending on the value of σ, the remaining columns provide the transmission time of the shortest path with respect to the lead time in the corresponding network. The symbol '-' means that the shortest path

 ISBN: 978-84-616-9216-3

obtained has a capacity greater than c_j if the network considered is $\mathcal{G}_j$ and therefore must not be taken into consideration. The letters 'NF' refer to non feasible, that is to say, there are no $s-t$ paths in the network considered. The two last rows in each part of the Table indicate the optimal solution and the CPU time in seconds invested in solving the problem.

It is worth pointing out that the CPU time needed to solve the problems is negligible (less than a second for all problems). As expected, when the value of σ increases, the actual number of arcs in each network decreases and the number of networks with no $s-t$ paths increases. Note also that, unlike the QPP, when σ increases the optimal solution of the EQPP is not necessarily an $s-t$ path with larger capacity, since the available energy at nodes limits the feasible $s-t$ paths.

Acknowledgements

This research work has been funded by the Gobierno de Aragón under grant E58 (FSE) and by UZ-Santander under grant UZ2012-CIE-07.

References

[1] H. I. Calvete, *The quickest path problem with interval lead times*, Computers and Operations Research **31** (2004) 383–395.

[2] H. I. Calvete and L. del Pozo, *The quickest path problem with batch constraints*, Operations Research Letters **31** (2003) 277–284.

[3] H. I. Calvete, L. del Pozo and J. A. Iranzo, *Algorithms for the quickest path problem and the reliable quickest path problem*, Computational Management Science **9** (2012) 255–272.

[4] H. I. Calvete, L. del Pozo and J. A. Iranzo, *The energy-constrained quickest path problem*, Preprint (2014).

[5] Y. L. Chen and Y. H. Chin, *The quickest path problem*, Computers and Operations Research **17** (1990) 153–161.

[6] J. C. N. Clímaco, M. M. B. Pascoal, J. M. F. Craveirinha and M. E. V. Captivo, *Internet packet routing: Application of a k-quickest path algorithm*, European J. of Operational Research **181** (2007) 1045–1054.

[7] H. W. Hamacher and S. A. Tjandra, *Mathematical modelling of evacuation problems: A state of the art*, In M. Schreckenberg and S. D. Sharma, editors, *Pedestrian and Evacuation Dynamics*, Springer, Berlin (2002) 227–266.

Table 1: $n = 1000$, $m = 10000$

$r = 10$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	-	-	-	NF
$\mathcal{G}_2$	147	897	1730	NF
$\mathcal{G}_3$	113	504	939	-
$\mathcal{G}_4$	110	456	841	1995
$\mathcal{G}_5$	111	433	790	1868
$\mathcal{G}_6$	-	-	-	NF
$\mathcal{G}_7$	-	-	-	NF
$\mathcal{G}_8$	127	408	720	NF
$\mathcal{G}_9$	148	367	618	NF
$\mathcal{G}_{10}$	NF	NF	NF	NF
Optimal solution	110	367	618	1868
CPU time (seconds)	0.01	0	0	0

$r = 20$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	188	1313	2563	NF
$\mathcal{G}_2$	178	1178	2289	NF
$\mathcal{G}_3$	-	-	-	4276
$\mathcal{G}_4$	-	-	-	-
$\mathcal{G}_5$	144	674	1262	3048
$\mathcal{G}_6$	150	650	1206	NF
$\mathcal{G}_7$	-	-	1107	NF
$\mathcal{G}_8$	153	562	1041	NF
$\mathcal{G}_9$	-	-	1030	NF
$\mathcal{G}_{10}$	-	510	-	NF
$\mathcal{G}_{11}$	-	-	-	NF
$\mathcal{G}_{12}$	-	478	823	NF
$\mathcal{G}_{13}$	-	501	834	NF
$\mathcal{G}_{14}$	-	-	838	NF
$\mathcal{G}_{15}$	149	414	845	NF
$\mathcal{G}_{16}$	193	412	NF	NF
$\mathcal{G}_{17}$	218	495	NF	NF
$\mathcal{G}_{18}$	NF	NF	NF	NF
$\mathcal{G}_{19}$	NF	NF	NF	NF
$\mathcal{G}_{20}$	NF	NF	NF	NF
Optimal solution	144	412	823	3048
CPU time (seconds)	0	0.01	0	0

 ISBN: 978-84-616-9216-3

Table 2: $n = 1000$, $m = 20000$

$r = 10$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	-	-	-	5050
$\mathcal{G}_2$	93	593	1149	2815
$\mathcal{G}_3$	112	541	1017	NF
$\mathcal{G}_4$	-	-	899	NF
$\mathcal{G}_5$	136	482	875	NF
$\mathcal{G}_6$	153	475	833	NF
$\mathcal{G}_7$	167	467	NF	NF
$\mathcal{G}_8$	192	473	NF	NF
$\mathcal{G}_9$	NF	NF	NF	NF
$\mathcal{G}_{10}$	NF	NF	NF	NF
Optimal solution	93	467	833	2815
CPU time (seconds)	0	0	0	0

$r = 20$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	-	NF	NF	NF
$\mathcal{G}_2$	-	5057	NF	NF
$\mathcal{G}_3$	-	-	-	NF
$\mathcal{G}_4$	-	-	-	-
$\mathcal{G}_5$	-	-	-	4239
$\mathcal{G}_6$	93	593	1149	2833
$\mathcal{G}_7$	-	NF	NF	NF
$\mathcal{G}_8$	132	NF	NF	NF
$\mathcal{G}_9$	NF	NF	NF	NF
$\mathcal{G}_{10}$	NF	NF	NF	NF
$\mathcal{G}_{11}$	NF	NF	NF	NF
$\mathcal{G}_{12}$	NF	NF	NF	NF
$\mathcal{G}_{13}$	NF	NF	NF	NF
$\mathcal{G}_{14}$	NF	NF	NF	NF
$\mathcal{G}_{15}$	NF	NF	NF	NF
$\mathcal{G}_{16}$	NF	NF	NF	NF
$\mathcal{G}_{17}$	NF	NF	NF	NF
$\mathcal{G}_{18}$	NF	NF	NF	NF
$\mathcal{G}_{19}$	NF	NF	NF	NF
$\mathcal{G}_{20}$	NF	NF	NF	NF
Optimal solution	93	593	1149	2833
CPU time (seconds)	0	0.01	0.01	0.01

 ISBN: 978-84-616-9216-3

Table 3: $n = 5000$, $m = 50000$

$r = 10$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	-	-	10057	25070
$\mathcal{G}_2$	-	-	-	-
$\mathcal{G}_3$	168	1293	2543	6297
$\mathcal{G}_4$	122	596	1122	NF
$\mathcal{G}_5$	134	563	1039	2469
$\mathcal{G}_6$	-	-	-	1971
$\mathcal{G}_7$	145	455	800	1894
$\mathcal{G}_8$	180	480	821	1918
$\mathcal{G}_9$	355	612	NF	NF
$\mathcal{G}_{10}$	NF	NF	NF	NF
Optimal solution	122	455	800	1894
CPU time (seconds)	0.02	0.01	0.01	0.01

$r = 20$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	543	5043	10057	25057
$\mathcal{G}_2$	-	-	-	-
$\mathcal{G}_3$	-	-	-	-
$\mathcal{G}_4$	159	1159	2270	5603
$\mathcal{G}_5$	-	-	-	NF
$\mathcal{G}_6$	128	821	1612	NF
$\mathcal{G}_7$	-	-	-	NF
$\mathcal{G}_8$	-	-	-	2600
$\mathcal{G}_9$	107	536	1012	2492
$\mathcal{G}_{10}$	-	-	-	2294
$\mathcal{G}_{11}$	114	460	845	2071
$\mathcal{G}_{12}$	-	-	857	2016
$\mathcal{G}_{13}$	-	-	831	1986
$\mathcal{G}_{14}$	133	443	801	1941
$\mathcal{G}_{15}$	153	453	847	NF
$\mathcal{G}_{16}$	209	466	755	NF
$\mathcal{G}_{17}$	NF	NF	NF	NF
$\mathcal{G}_{18}$	NF	NF	NF	NF
$\mathcal{G}_{19}$	NF	NF	NF	NF
$\mathcal{G}_{20}$	NF	NF	NF	NF
Optimal solution	107	443	755	1941
CPU time (seconds)	0.03	0.02	0.01	0.02

 ISBN: 978-84-616-9216-3

Table 4: $n = 5000$, $m = 100000$

$r = 10$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	-	-	20052	NF
$\mathcal{G}_2$	-	-	-	12543
$\mathcal{G}_3$	-	-	-	-
$\mathcal{G}_4$	151	1151	2262	5595
$\mathcal{G}_5$	106	606	1162	2828
$\mathcal{G}_6$	-	-	-	NF
$\mathcal{G}_7$	-	-	-	NF
$\mathcal{G}_8$	102	477	894	NF
$\mathcal{G}_9$	223	428	NF	NF
$\mathcal{G}_{10}$	NF	NF	NF	NF
Optimal solution	102	428	894	2828
CPU time (seconds)	0.02	0.03	0.03	0.03

$r = 20$	σ			
	1000	5000	20000	50000
$\mathcal{G}_1$	-	-	20057	NF
$\mathcal{G}_2$	-	-	-	12544
$\mathcal{G}_3$	-	-	-	-
$\mathcal{G}_4$	206	1706	3373	-
$\mathcal{G}_5$	157	1157	2268	5601
$\mathcal{G}_6$	139	958	-	4595
$\mathcal{G}_7$	-	-	-	2828
$\mathcal{G}_8$	-	-	-	-
$\mathcal{G}_9$	-	-	-	-
$\mathcal{G}_{10}$	91	466	883	2133
$\mathcal{G}_{11}$	103	463	863	2063
$\mathcal{G}_{12}$	105	451	836	1992
$\mathcal{G}_{13}$	113	446	816	1933
$\mathcal{G}_{14}$	-	-	-	-
$\mathcal{G}_{15}$	122	412	737	1709
$\mathcal{G}_{16}$	139	370	626	1396
$\mathcal{G}_{17}$	149	364	602	1374
$\mathcal{G}_{18}$	186	391	635	1406
$\mathcal{G}_{19}$	NF	NF	NF	NF
$\mathcal{G}_{20}$	NF	NF	NF	NF
Optimal solution	91	364	602	1374
CPU time (seconds)	0.05	0.06	0.05	0.04

 ISBN: 978-84-616-9216-3

[8] D. Klingman, A. Napier and J. Stutz, *Netgen: A program for generating large scale capacitated assignment, transportation, and minimum cost flow network problems*, Management Science **20** (1974) 814–821.

[9] E. Q. V. Martins and J. L. E. Santos, *An algorithm for the quickest path problem*, Operations Research Letters **20** (1997) 195–198.

[10] M. H. Moore, *On the fastest route for convoy-type traffic in flowrate-constrained networks*, Transportation Science **10** (1976) 113–124.

[11] C. K. Park, S. Lee and S. Park, *A label-setting algorithm for finding a quickest path*, Computers and Operations Research **31** (2004) 2405–2418.

[12] M. M. B. Pascoal, M. E. V. Captivo and J. C. N. Clímaco, *A comprehensive survey on the quickest path problem*, Annals of Operations Research **147** (2006) 5–21.

[13] J. B. Rosen, S. Z. Sun and G. L. Xue, *Algorithms for the quickest path problem and the enumeration of quickest paths*, Computers and Operations Research **18** (1991) 579–584.

[14] S. Ruzika and M. Thiemann, *Min-max quickest path problems*, Networks **60** (2012) 253–258.

[15] A. Sedeño-Noda and J. D. González-Barrera, *Fast and fine quickest path algorithm*, European J. of Operational Research, DOI: http://dx.doi.org/10.1016/j.ejor.2014.04.028, (2014).

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Pricing fixed-rate mortgages under jump-diffusion models for the house value

María del Carmen Calvo-Garrido[1] **and Carlos Vázquez**[1]

[1] *Department of Mathematics, University of A Coruna*

emails: mcalvog@udc.es, carlosv@udc.es

Abstract

In the pricing of fixed rate mortgages with prepayment and default options we introduce jump-diffusion models for the house price evolution. These models take into account sudden changes in the price (jumps) during bubbles and crisis situations in real state markets. After posing the models based on partial integro-differential equations (PIDE) problems for the contract, insurance and coinsurance, we propose appropriate numerical methods to solve them.

Key words: Fixed-rate mortgages, jump-diffusion models, complementarity problems, numerical methods

1 Introduction

A mortgage is a financial product in which the borrower obtains funds by using a risky asset as collateral, usually a house. The loan is reimbursed through monthly payments until the cancelation of the debt at maturity date. Thus, the mortgage value is understood as the discounted value of the future monthly payments (without including a possible insurance on the loan by the lender) and the underlying stochastic factors are the house price and the interest rate. In this work we follow [11, 4], where early prepayment is allowed at any time and default can occur at any monthly payment date. In both previous papers a lognormal process is assumed for the house price evolution so that the this value evolves continuously. However, in certain situations, such as during the relatively recent bubble or crisis phenomena in real state markets, the consideration of the standard lognormal process is no longer so realistic. Thus, it becomes necessary to consider jump-diffusion models to account with sudden changes in the value of the house and this is the main innovative point of the present work.

In the forthcoming sections we briefly describe the pricing model under consideration as well as the mortgage contract related aspects. Next, we consider the numerical solution techniques and finally we present some numerical results allowing to compare the case without jumps and two different jump-diffusion models here proposed.

2 Mathematical modelling

In order to model the evolution of the house value at time t, H_t, we consider the following stochastic differential equation (SDE):

$$dH_t = (\mu - \delta)H_t\,dt \,+\, \sigma_H H_t\,dX_t^H \,+\, d\left(\sum_{i=1}^{N_t} Y_i\right), \tag{1}$$

where μ, δ and σ_H denote house appreciation average rate, the dividend yield provided by (hiring or using) the house and the house price volatility, respectively, while dX_t^H represents a Wiener process for the house price. Moreover, in the jump part of the model $(N_t)_{t\geq 0}$ denotes a Poisson process with parameter $\tilde{\lambda}$ and (Y_i) represents a sequence of square integrable, independent and identically distributed random variables, so that X_t^H, N_t and (Y_i) are independent. In order to complete the model definition, we specify the distribution of jump sizes by using either a Merton [9] or a Kou [8] model. More precisely, under Merton model (Y_i) are taken from the lognormal distribution $LN(\mu_j, \gamma_j^2)$, with density function

$$\nu_m(y) = \frac{1}{y\gamma_j\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu_j)^2}{2\gamma_j^2}\right), \tag{2}$$

where μ_j and γ_j are the mean and the standard deviation of the jump size, respectively, whereas under Kou model (Y_i) follows a distribution with the log-double-exponential density

$$\nu_k(y) = \begin{cases} q\alpha_2 y^{\alpha_2 - 1}, & y < 1 \\ p\alpha_1 y^{-\alpha_1 - 1}, & y \geq 1, \end{cases} \tag{3}$$

where p, q, α_1 and α_2 are positive constants such that $p + q = 1$ and $\alpha_1 > 1$. Note that p and q represent the probabilities of upward and downward jumps, respectively.

Under a risk neutral probability measure, we can obtain the equivalent SDE:

$$dH_t = (r_t - \delta)H_t dt + \sigma_H H_t dX_t^H \,+\, d\left(\sum_{i=1}^{N_t} Y_i\right). \tag{4}$$

Additionally, we assume that the interest rate follows the CIR following process [4]:

$$dr_t = \kappa(\theta - r_t)dt + \sigma_r\sqrt{r_t}dX_t^r, \tag{5}$$

 ISBN: 978-84-616-9216-3

where κ denotes the speed of mean reversion to the long term rate θ and σ_r denotes a volatility parameter. Wiener processes X_t^H and X_t^r could be correlated with correlation coefficient ρ (i.e. $dX_t^H dX_t^r = \rho dt$) to incorporate possible correlation between interest rate and house price.

2.1 Partial integral differential equation (PIDE) formulation

By using a dynamic hedging technique, in the case without jumps a PDE model for pricing any asset depending on house price and interest rate is posed in [4]. In the here treated jump-diffusion models for the house price, if we assume that the value of any asset depending on house price and interest rate is given by $F_t = F(t, H_t, r_t)$, then the function F satisfies the following partial integral differential equation (PIDE):

$$\partial_t F + \frac{1}{2}\sigma_H^2 H^2 \partial_{HH} F + \rho\sigma_H\sigma_r H\sqrt{r}\partial_{Hr}F + \frac{1}{2}\sigma_r^2 r\partial_{rr}F + (r-\delta)H\partial_H F + \kappa(\theta - r)\partial_r F - rF + \int_0^\infty \tilde{\lambda}\left[F(t,Hy,r) - F(t,H,r) - H(y-1)\partial_H F(t,H,r)\right]\nu(y)dy = 0, \quad (6)$$

where subindexes in ∂ indicate partial derivatives and $\nu(y) = \nu_m(y)$ for Merton model, whereas $\nu(y) = \nu_k(y)$ for Kou one. Since ν is a probability density function then

$$\int_0^\infty \nu(y)dy = 1.$$

Moreover, we can compute the expectations for Merton and Kou models

$$E_m[Y_i] = \int_0^\infty y\nu_m(y)dy = e^{\mu_j + \gamma_j^2/2}, \quad E_k[Y_i] = \int_0^\infty y\nu_k(y)dy = \frac{p\alpha_1}{\alpha_1 - 1} + \frac{q\alpha_2}{\alpha_2 + 1}.$$

Therefore, the PIDE (6) can be written in the form

$$\partial_t F + \frac{1}{2}\sigma_H^2 H^2 \partial_{HH} F + \rho\sigma_H\sigma_r H\sqrt{r}\partial_{Hr}F + \frac{1}{2}\sigma_r^2 r\partial_{rr}F + (r-\delta-\tilde{\lambda}\tilde{\kappa})H\partial_H F + \kappa(\theta - r)\partial_r F - (r+\tilde{\lambda})F + \tilde{\lambda}\int_0^\infty F(t,Hy,r)\nu(y)dy = 0, \quad (7)$$

where

$$\tilde{\kappa} = e^{\mu_j + \gamma_j^2/2} - 1 \quad or \quad \tilde{\kappa} = \frac{p\alpha_1}{\alpha_1 - 1} + \frac{q\alpha_2}{\alpha_2 + 1} - 1$$

for Merton or Kou models, respectively.

Note that with respect to the PDE model in [4], there is an additional integral term in the equation due to the presence of jumps. This term makes the PIDE more difficult to solve than the corresponding PDE.

 ISBN: 978-84-616-9216-3

2.2 Mortgage contract

Following the same notation as in [4], the equal monthly payments dates are denoted by T_m, $m = 1, ..., M$, where M is the number of months. Assuming $T_0 = 0$, let $\Delta T_m = T_m - T_{m-1}$ be the duration of month m, c the fixed contract rate and $P(0)$ the initial loan (i.e. the principal at $t = T_0 = 0$), then the fixed mortgage payment (MP) is given by:

$$MP = \frac{(c/12)(1 + c/12)^M P(0)}{(1 + c/12)^M - 1}. \tag{8}$$

For $m = 1, ..., M$, the unpaid loan just after the $(m-1)$th payment date is

$$P(m-1) = \frac{((1 + c/12)^M - (1 + c/12)^{m-1})P(0)}{(1 + c/12)^M - 1}, \tag{9}$$

If $t_m = t - T_{m-1}$ denotes the time elapsed at month m (which starts at $t = T_{m-1}$), let $\tau_m = \Delta T_m - t_m$ be the time until T_m. This change of variable transforms equation (7) into another one associated with an initial value problem. More precisely, the mortgage value to the lender during month m, $V(\tau_m, H, r)$, satisfies the PIDE

$$\begin{aligned} -\partial_{\tau_m} F + \frac{1}{2}\sigma_H^2 H^2 \partial_{HH} F + \rho\sigma_H\sigma_r H\sqrt{r}\partial_{Hr} F + \frac{1}{2}\sigma_r^2 r \partial_{rr} F + \\ (r - \delta - \tilde{\lambda}\tilde{\kappa})H\partial_H F + \kappa(\theta - r)\partial_r F - (r + \tilde{\lambda})F + \tilde{\lambda}\int_0^\infty F(\tau_m, Hy, r)\nu(y)dy &= 0, \end{aligned} \tag{10}$$

for $0 \le \tau_m \le \Delta T_m$, $0 \le H < \infty$, $0 \le r < \infty$. We clarify a certain abuse of notation: if $\overline{F}$ denotes the solution of (7) and F the solution of (10) then $F(\tau_m, H, r) = \overline{F}(T_m - \tau_m, H, r)$.

Next, we take into account the prepayment and default options. The option to default only happens at payment dates when the borrower does not pay the amount MP. The option to prepay can be exercised at any time during the life of the loan. In the case of prepayment the borrower fully amortizes the mortgage at time τ_m by paying the following amount (which includes the total remaining debt plus an early termination penalty):

$$TD(\tau_m) = (1 + \Psi)(1 + c(\Delta T_m - \tau_m))P(m-1), \tag{11}$$

where Ψ denotes the prepayment penalty factor.

The mortgage pricing problem starts from the value of the mortgage at maturity ($t = T_M$), just before the last payment, given by:

$$V(\tau_M = 0, H, r) = \min(MP, H), \tag{12}$$

while at the other payment dates ($1 \le m \le M - 1$), it is given by

$$V(\tau_m = 0, H, r) = \min(V(\tau_{m+1} = \Delta T_{m+1}, H, r) + MP, H). \tag{13}$$

 ISBN: 978-84-616-9216-3

If the borrower defaults, which occurs when the mortgage value is equal to the house value, the lender will lose the promised future payments, unless an insurance against default covering a fraction of the loss has been taken. This insurance contract has no value for the borrower, as it is part of the lender's portfolio [11]. In order to obtain the value of the insurance to the lender, denoted by $I(\tau_m, H, r)$, we must solve equation (10) with suitable payment date conditions. We assume that in case of default the insurer accepts to pay a fraction γ of the currently unpaid balance to up to a maximum indemnity, Γ. Therefore, depending if default occurs or not, the insurance value at the maturity of the loan is

$$I(\tau_M = 0, H, r) = \begin{cases} \min(\gamma(MP - H), \Gamma) & \text{(Default)} \\ \\ 0 & \text{(No default)} \end{cases} \tag{14}$$

At earlier payment dates ($1 \leq m \leq M-1$), the value of the insurance is

$$I(\tau_m = 0, H, r) = \begin{cases} \min(\gamma[TD(\tau_m = 0) - H], \Gamma) & \text{(Default)} \\ \\ I(\tau_{m+1} = \Delta T_{m+1}, H, r) & \text{(No default)} \end{cases} \tag{15}$$

The fraction of the potential loss not covered by the insurance is the coinsurance. At each payment date, the coinsurance is the difference between the values of the potential loss and the insurance coverage. In order to price the coinsurance, $CI(\tau_m, H, r)$, equation (10) must be solved again with suitable coinsurance conditions. At maturity, we consider

$$CI(\tau_M = 0, H, r) = \begin{cases} \max((1-\gamma)(MP - H), (MP - H) - \Gamma) & \text{(Default)} \\ \\ 0 & \text{(No default)} \end{cases} \tag{16}$$

while at earlier payment dates ($1 \leq m \leq M-1$), we consider

$$CI(\tau_m = 0, H, r) = \begin{cases} \max((1-\gamma)[TD(\tau_m = 0) - H], [TD(\tau_m = 0) - H] - \Gamma) & \text{(Default)} \\ \\ CI(\tau_{m+1} = \Delta T_{m+1}, H, r) & \text{(No default)} \end{cases} \tag{17}$$

At origination, the equilibrium condition explained in [4] needs to be satisfied in order to avoid arbitrage. Formally,

$$V(\tau_1 = \Delta T_1, H_{initial}, r_{initial}; \Psi, c) + I(\tau_1 = \Delta T_1, H_{initial}, r_{initial}; \Psi, c) = (1 - \xi)P(0). \tag{18}$$

The contract rate is adjusted by using the same iterative process as in [4].

2.3 The free boundary problem under jump-diffusion models

The option to prepay the loan at any time gives rise to a free boundary problem, in which not only the mortgage price is obtained but also the regions where it is optimal to fully amortize

the loan or not. Both regions are separated by a free boundary (optimal prepayment boundary). If we consider the following nonlocal linear operator:

$$\begin{aligned}\mathcal{L}_j V &= \partial_{\tau_m} V - \frac{1}{2}\sigma_H^2 H^2 \partial_{HH} V - \rho\sigma_H\sigma_r H\sqrt{r}\partial_{Hr}V - \frac{1}{2}\sigma_r^2 r\partial_{rr}V - \\ &\quad (r-\delta-\tilde{\lambda}\tilde{\kappa})H\partial_H V - \kappa(\theta-r)\partial_r V + (r+\tilde{\lambda})V - \tilde{\lambda}\int_0^\infty V(\tau_m, Hy, r)\nu(y)dy \quad (19)\end{aligned}$$

then the free boundary problem can be posed in terms of the linear complementarity one:

$$\mathcal{L}_j V \leq 0, \quad (TD(\tau_m) - V(\tau_m, H, r)) \geq 0, \quad (\mathcal{L}_j V)(TD(\tau_m) - V(\tau_m, H, r)) = 0. \quad (20)$$

In the region $V = TD$ it is optimal for the borrower to prepay, otherwise $\mathcal{L}_j V = 0$ and we are inside the region where we continue to pay the loan without prepayment.

3 Numerical solution

The PIDE is initially posed on an unbounded domain, so that we approximate it by a bounded domain formulation and we impose boundary conditions. Note that the domain of integration in the integral term also needs to be localized. We introduce the following changes of variables and notation:

$$x_1 = \frac{H}{H_\infty}, \quad x_2 = \frac{r}{r_\infty}, \quad \bar{x} = \log x_1, \quad \eta = log(y) \quad (21)$$

where both H_∞ and r_∞ are sufficiently large suitably chosen real numbers. Let $\Omega = (0, x_1^\infty) \times (0, x_2^\infty)$, with $x_1^\infty = x_2^\infty = 1$. Then, let us denote the Lipschitz boundary by $\Gamma = \partial\Omega$ such that $\Gamma = \bigcup_{i=1}^2 (\Gamma_i^- \cup \Gamma_i^+)$, where:

$$\Gamma_i^- = \{(x_1, x_2) \in \Gamma \mid x_i = 0\}, \; \Gamma_i^+ = \{(x_1, x_2) \in \Gamma \mid x_i = x_i^\infty\}, \; i = 1, 2.$$

Next, taking into account the new variables we write the equation (10) in divergence form in the bounded domain. As in [11], we consider the case $\rho = 0$. Thus, the initial-boundary value problem for the insurance and coinsurance can be written in the form:
Find $J : [0, \Delta T_m] \times \Omega \to \mathbb{R}$ such that

$$\frac{\partial J}{\partial \tau_m} + \vec{v}\cdot\nabla J - Div(A\nabla J) + lJ - \tilde{\lambda}\int_{\eta min}^{\eta max} \bar{J}(\tau_m, \bar{x}_1 + \eta, x_2)\bar{\nu}(\eta)d\eta = f \qquad \text{in } (0, \Delta T_m)\times\Omega \quad (22)$$

$$\frac{\partial J}{\partial x_1} = g_1 \qquad \text{on } (0, \Delta T_m)\times\Gamma_1^+ \quad (23)$$

$$\frac{\partial J}{\partial x_2} = g_2 \qquad \text{on } (0, \Delta T_m)\times\Gamma_2^+ \quad (24)$$

where $J = I, CI$ and the appropriate initial condition for each month is given by the equations (14) and (15) when we are pricing the insurance and by the equations (16) and (17) in the case of valuing the coinsurance.

 ISBN: 978-84-616-9216-3

Furthermore, for the complementarity problem associated with the mortgage value during month m, we can pose the following mixed formulation:

Find $V : [0, \Delta T_m] \times \Omega \to \mathbb{R}$ satisfying the partial differential equation

$$\frac{\partial V}{\partial \tau_m} + \vec{v} \cdot \nabla V - Div(A \nabla V) + lV - \tilde{\lambda} \int_{\eta min}^{\eta max} \bar{V}(\tau_m, \bar{x}_1 + \eta, x_2) \bar{\nu}(\eta) d\eta + P = 0, \tag{25}$$

the complementarity conditions

$$V \leq TD, \quad P \geq 0, \quad P(TD - V) = 0, \tag{26}$$

the boundary conditions

$$\frac{\partial V}{\partial x_1} = 0 \qquad \text{on } (0, \Delta T_m) \times \Gamma_1^+ , \tag{27}$$

$$\frac{\partial V}{\partial x_2} = 0 \qquad \text{on } (0, \Delta T_m) \times \Gamma_2^+ , \tag{28}$$

and the initial condition for each month, given by the equations (12) or (13).

For both problems, the involved data is defined as follows

$$A = \begin{pmatrix} \frac{1}{2}\sigma_H^2 x_1^2 & 0 \\ 0 & \frac{1}{2}\sigma_r^2 \frac{x_2}{r_\infty} \end{pmatrix}, \quad \vec{v} = \begin{pmatrix} (\sigma_H^2 - x_2 r_\infty + \delta + \tilde{\lambda}\tilde{\kappa}) x_1 \\ (\frac{1}{2}\sigma_r^2 - \kappa(\theta - x_2 r_\infty))/r_\infty \end{pmatrix}, \quad l = x_2 r_\infty + \tilde{\lambda}. \tag{29}$$

Remark 3.1 *Note that the differential term of the PIDE is computed in the domain* $[0, x_1^\infty] \times [0, x_2^\infty]$*, using the discrete grid:* $0 = x_{1_0}, x_{1_1}, \cdots, x_{1_q} = x_1^\infty$*. Since* $log(x_{1_0}) = -\infty$*, we choose* $\eta min = log(x_{1_1})$ *and* $\eta max = log(x_{1_q})$ *as it is proposed in [6].*

Under Merton model, the function $\bar{\nu}$ is given by

$$\bar{\nu}(\eta) = \bar{\nu}_m(\eta) = \frac{1}{\gamma_j \sqrt{2\pi}} \exp\left(-\frac{(\eta - \mu_j)^2}{2\gamma_j^2} \right), \tag{30}$$

whereas under Kou model

$$\bar{\nu}(\eta) = \bar{\nu}_k(\eta) = \begin{cases} q\alpha_2 e^{\alpha_2 \eta}, & \eta < 0 \\ p\alpha_1 e^{-\alpha_1 \eta}, & \eta \geq 0. \end{cases} \tag{31}$$

Once the localization procedure has been carried out, we consider a Lagrange-Galerkin discretization based on a Crank-Nicolson scheme introduced in [2, 3]. Thus, we define the characteristics curve through $\mathbf{x} = (x_1, x_2)$ at time $\bar{\tau}_m$, $X(\mathbf{x}, \bar{\tau}_m; s)$, which satisfies:

$$\frac{\partial}{\partial s} X(\mathbf{x}, \bar{\tau}_m; s) = \vec{v}(X(\mathbf{x}, \bar{\tau}_m; s)), \; X(\mathbf{x}, \bar{\tau}_m; \bar{\tau}_m) = \mathbf{x}. \tag{32}$$

For $N > 1$ let us consider the time step $\Delta\tau_m = \Delta T_m/N$ and the time mesh points $\tau_m^n = n\Delta\tau_m$, $n = 0, \frac{1}{2}, 1, \frac{3}{2}, \ldots, N$. The material derivative approximation by characteristics method is given by:

$$\frac{DF}{D\tau_m} = \frac{F^{n+1} - F^n \circ X^n}{\Delta\tau_m},$$

where $F = CI$, I , V and $X^n(\mathbf{x}) := X(\mathbf{x}, \tau_m^{n+1}; \tau_m^n)$. In view of the expression of the velocity field the components of $X^n(\mathbf{x})$ can be analytically computed:

$$\begin{aligned} X_1^n(\mathbf{x}) &= x_1 \exp\left(-\left(\sigma_H^2 + \delta + \frac{\sigma_r^2}{2\kappa} - \theta + \tilde{\lambda}\tilde{\kappa}\right)\Delta\tau_m\right) \times \\ &\quad \exp\left(\left(\frac{-x_2 r_\infty}{\kappa} - \frac{\sigma_r^2}{2\kappa^2} + \frac{\theta}{\kappa}\right)(\exp(-\kappa\Delta\tau_m) - 1)\right) \\ X_2^n(\mathbf{x}) &= \left(-\frac{\sigma_r^2}{2\kappa r_\infty} + \frac{\theta}{r_\infty}\right)(1 - \exp(-\kappa\Delta\tau_m)) + x_2 \exp(-\kappa\Delta\tau_m) \end{aligned}$$

Next, we consider a Crank-Nicolson scheme around $\left(X(\mathbf{x}, \tau_m^{n+1}; \tau_m), \tau_m\right)$ for $\tau_m = \tau_m^{n+\frac{1}{2}}$. So, for $n = 0, \ldots, N-1$, the time discretized equation for $F = I$, CI, V and $P = 0$ can be written as follows:

Find F^{n+1} such that:

$$\begin{aligned} &\frac{F^{n+1}(\mathbf{x}) - F^n(X^n(\mathbf{x}))}{\Delta\tau_m} - \frac{1}{2}Div(A\nabla F^{n+1})(\mathbf{x}) - \frac{1}{2}Div(A\nabla F^n)(X^n(\mathbf{x})) + \\ &\quad \frac{1}{2}(l\, F^{n+1})(\mathbf{x}) + \frac{1}{2}(l\, F^n)(X^n(\mathbf{x})) - \tilde{\lambda}\int_{\eta min}^{\eta max} \bar{F}^n(\bar{x}_1 + \eta, x_2)\bar{\nu}(\eta)d\eta = 0, \quad (33) \end{aligned}$$

where $\bar{F}^n(\bar{x}_1 + \eta, x_2) = F^n(e^{\bar{x}_1+\eta}, x_2)$. Note that the integral term is evaluated at the previous time step, thus avoiding the presence of a full matrix in the linear systems associated to the fully discretized problem [5, 10].

Next, we can write a variational formulation for the semi-discretized problems and use piecewise quadratic Lagrange finite elements for spatial discretization. In order to deal with the nonlinearities in the free boundary problem associated to prepayment option, we implement the ALAS algorithm proposed in [7] and explained in detail in [4] for the case without jumps in the house price.

In order to approximate the integral term that appears in the PIDE due to the presence of jumps we use a suitable numerical integration procedure. More precisely, we use the classical composite trapezoidal rule with $m+1$ points in the following way:

$$\begin{aligned} &\int_{\eta min}^{\eta max} \bar{F}^n(\bar{x}_1 + \eta, x_2)\bar{\nu}(\eta)d\eta \approx \\ &\frac{h}{2}\left[\bar{F}^n(\bar{x}_1 + \eta min, x_2)\bar{\nu}(\eta min) + \bar{F}^n(\bar{x}_1 + \eta max, x_2)\bar{\nu}(\eta max) + 2\sum_{j=1}^{m-1} \bar{F}^n(\bar{x}_1 + k_j, x_2)\bar{\nu}(k_j)\right], \end{aligned}$$

 ISBN: 978-84-616-9216-3

where $k_j = \eta min + jh$ for $j = 1, ..., m-1$ and $h = \frac{\eta max - \eta min}{m}$.

4 Numerical results

In order to solve the fixed rate mortgage valuation problem, we need to specify a set of parameters related to the stochastic models, contract characteristics and insurance. Most of them are based on the existent literature (see [1] and [11], for example) and are shown in Table 1. Moreover, concerning the numerical methods employed to solve the problem, we consider the parameters collected in Table 2. In order to compare the results obtained with Merton and Kou models we need that the density functions of the normal distribution and of the double-exponential distribution match. For this purpose, we consider the parameters involved in the jump-diffusion models which are proposed in [6]

House price and interest rate models data	
Steady state spot rate, θ	10 %
Speed of reversion, κ	25 %
House service flow, δ	7.5%
House price volatility, σ_H	5%
Interest rate volatility, σ_r	5%
Parameter of Poisson process, $\tilde{\lambda}$	0.1
Mean of jump size (Merton), μ_j	-0.1
Standard deviations of jump size (Merton), γ_j	0.45
Probability of upward jump (Kou), p	0.3445
Parameter (Kou), α_1	3.0465
Parameter (Kou), α_2	3.0775
Contract specifications	
Loan maturity (years)	15
Initial value of the house, $H_{initial}$	100000
Spot interest rate, $r_{initial}$	8 %
Ratio of the loan to value	95 %
Initial estimate for contract rate, c_0	10%
Prepayment penalty, Ψ	5%
Arrangement fee, ξ	0%
Insurance	
Guaranteed fraction of total loss, γ	80%
Cap, Γ	$20\%H_{initial}$

Table 1: Fixed parameters in the mortgage valuation model

 ISBN: 978-84-616-9216-3

In Table 3 we show a comparison between the contract values without and with jumps for the house price. In the presence of jumps we take into account Merton and Kou models. As expected, in the absence of jumps the value of the contract is higher than with jumps whereas the value of the insurance and the coinsurance are lower. Note that the presence of jumps increases uncertainty in the house price, thus depreciating the mortgage price.

We also note that the prepayment region is located in the part of the domain with lower rates and higher house prices as in the case without jumps [4], which results reasonable from the financial point of view: in this part it is better to fully prepay the loan and refinance at lower market interest rates if necessary.

Computational domain	
H_∞	200000
r_∞	40 %
Finite elements mesh data	
Number of elements	576
Number of nodes	2401
Time discretization	
Time steps per month	30
ALAS algorithm	
Parameter β	10000

Table 2: Numerical resolution and jump-diffusion model parameters

	Contract rate c	Contract value V	Insurance I	Coinsurance CI
Without jumps	9.0839%	94549	449	112
Merton model	14.4301%	91730	3270	2402
Kou model	14.2355%	92090	2910	2092

Table 3: Comparison of the values obtained without and with jumps for the house value

Acknowledgements

This paper has been partially funded by MCINN (Project MTM2010–21135–C02-01) and by Xunta de Galicia (Ayuda CN2011/004 cofunded with FEDER funds).

 ISBN: 978-84-616-9216-3

References

[1] J. A. Azevedo-Pereira, D. P. Newton and D. A. Paxson, *UK Fixed Rate Repayment Mortgage and Mortgage Indemnity Valuation*, Real Estate Economics, **30** (2002), 185-211.

[2] A. Bermúdez, M. R. Nogueiras and C. Vázquez, *Numerical analysis of convection-diffusion-reaction problems with higher order characteristics finite elements. Part I: Time discretization*, SIAM Journal on Numerical Analysis, **44** (2006), 1829-1853.

[3] A. Bermúdez, M. R. Nogueiras and C. Vázquez, *Numerical analysis of convection-diffusion-reaction problems with higher order characteristics finite elements. Part II: Fully discretized scheme and quadrature formulas*, SIAM Journal on Numerical Analysis, **44** (2006), 1854-1876.

[4] M. C. Calvo-Garrido, C. Vázquez, *A new numerical method for pricing Fixed-Rate Mortgages with prepayment and default options*, accepted for publication in International Journal of Computer Mathematics (2013).

[5] R. Cont and P. Tankov, *Financial Modelling With Jump Processes*, Chapman & Hall/CRC Financial Mathematics Series, 2004.

[6] Y. D'Halluin, P. A. Forsyth and K.R. Vetzal, *Robust numerical methods for contingent claims under jump diffusion processes*, IMA Journal of Numerical Analysis, **25** (2005), 87-112.

[7] T. Kärkkäinen, K. Kunisch, P. Tarvainen, *Augmented Lagrangian Active Set methods for obstacle problems*, Journal of Optimization Theory and Applications, **19**, no. 3 (2003), 499-533.

[8] S. G. Kou, *A jump-diffusion model for option pricing*, Management Science, **48** (2002), 1086-1101.

[9] R. C. Merton, *Option pricing when underlying stock returns are discontinuous*, J. Finan. Econ., **3** (1976), 125-144.

[10] S. Salmi, J. Toivanen, *An iterative method for pricing American options under jump-diffusion models*, Applied Numerical Mathematics, **61** (2011), 821-831.

[11] N. J. Sharp, D. P. Newton and P. W. Duck, *An improved fixed-rate mortgage valuation methodology with interacting prepayment and default options*, Journal of Real Estate Finance and Economics, **19** (2008), 49-67.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

On the effective thermal conductivity for a transversely isotropic two phase composite material

Carmen Calvo-Jurado[1] **and William J. Parnell**[2]

[1] *Departamento de Matemticas, Escuela Politécnica, Avda. de la Universidad s/n, 10003, Cáceres, Spain, Universidad de Extremadura*

[2] *School of Mathematics, Alan Turing Building, Manchester* M13 9PL*, UK, University of Manchester*

emails: ccalvo@unex.es, william.parnell@manchester.ac.uk

Abstract

This paper is concerned with the estimation of the effective thermal conductivity of a transversely two phase composite. We derive a straightforward way to construct the Hashin-Shtrikman bounds from first principles in conductivity and taking into account the microstructure of the problem that it is governed by spheroidal statistics. That the shape of the inclusions and their distribution can be specified independently is of great utility in composite design. This case covers a multitude of composites used in applications by taking various limits of the spheroid, including both layered media and long unidirectional composites. Of specific interest is the fact that the corresponding Hill tensors can be derived analytically. We exhibit the implementation of the constructions with several examples.

1 Extended abstract

Let us consider a two phase composite material occupying a domain $\Omega \subset \mathbb{R}^N$ and $\varepsilon > 0$ a parameter taking its values in a sequence which goes to zero. The composite is described by the properties of its two phases Ω_ε^1 and Ω_ε^2, with thermal conductivity tensors $\mathbf{K}^1$, $\mathbf{K}^2$ and volume fractions ϕ^1 and ϕ^2 respectively. The microstructure of the mixture whose size is represented by ε, is determined by the geometric arrangement of the phases Ω_ε^1 and Ω_ε^2 satisfying $\Omega_\varepsilon^1 \cup \Omega_\varepsilon^2 = \Omega$, $\Omega_\varepsilon^1 \cap \Omega_\varepsilon^2 = \emptyset$. The problem governing the steady state temperature T_ε is given by the following linear elliptic problem with Dirichlet boundary conditions

 ISBN: 978-84-616-9216-3

$$\begin{cases} -div\,(\mathbf{K}_\varepsilon(x)\nabla T_\varepsilon) = f & \text{in } \Omega \\ T_\varepsilon = 0 & \text{on } \partial\Omega \end{cases} \tag{1}$$

where we have denoted by f the internal source term and by $\mathbf{K}_\varepsilon$ the thermal conductivity of the composite, that satisfies

$$\mathbf{K}_\varepsilon(x) = \begin{cases} \mathbf{K}^1(x) & \text{if } x \in \Omega_\varepsilon^1 \\ \mathbf{K}^2(x) & \text{if } x \in \Omega_\varepsilon^2 \end{cases}, \quad \mathbf{K}_\varepsilon(x) = \mathbf{K}^1(x)\chi_{\Omega_\varepsilon^1}(x) + \mathbf{K}^2(x)\chi_{\Omega_\varepsilon^2}(x)$$

We will assume that the two phase materials are transversely isotropic, i.e., there exist κ_1^r, κ_2^r, such that $K_{ij}^r = \kappa_1^r\Theta_{ij} + \kappa_2^r\delta_{i3}\delta_{j3}$, with $\Theta_{ij} = \delta_{ij} - \delta_{i3}\delta_{j3}$, $r = 1, 2$.

Our aim in this paper is to estimate the effective properties of the composite for a sufficiently small ε, i.e. when the microstructure is getting finer. However, a detailed knowledge of this kind o problem is too difficult and it is usual characterize the composite by its macroscopic effective properties, represented by a two order conductivity tensor $\mathbf{K}^*$. There are different methods to treat this problem. One of them is the homogenization theory, that passing to the limit as ε tends to zero, gives an homogenized limit problem that has the advantage that the matrix coefficients are constants, and then permits one determine the effective or average properties of the mixture. Indeed, there are some well known results that characterize its eigenvalues due to L. Tartar ([9], [5]) for the N-dimensional case, or in a two dimensional setting, to Lurie & Cherkaev ([2]).

Besides the homogenization techniques, the use of expressions defining bounds for physical properties of a mixture has been studied extensively by several authors ([1], [4], [7],[8]), because give the possible range of variation for such properties. In the conductivity setting, the Maxwell principle for the conductivity of a host material containing a suspension of spheres, is the most known work ([3]). When the only information about the microstructure are the volume fraction ϕ^r and the conductivity tensor $\mathbf{K}^r$ regarding to the r-th phase, $r = 0, 1, \ldots, n$, the effective conductive tensor $\mathbf{K}^*$ for a microstructure of arbitrary symmetry (isotropic or anisotropic) can be estimated by the Voigt $\mathbf{K}^V$ ([8]) and the Reuss $\mathbf{K}^R$ ([7]) bounds, as follows

$$K_{ij}^R \leq K_{ij}^* \leq K_{ij}^V, \qquad (K^R)_{ij}^{-1} = \sum_{r=0}^{n} \phi^r (K^r)_{ij}^{-1}, \qquad K_{ij}^V = \sum_{r=0}^{n} \phi^r K_{ij}^r.$$

The Voigt-Reuss bounds only depend on the phase volume fraction and are independent of any characteristic of the symmetry of the microstructure. Therefore, they are usually two wide to be of predictive interest. Using a variational principle, better results were obtained by Hashin & Shtrikman ([1]), who provided the tightest possible range of variation for the property of interest without information about the distribution of the phases. In the case of an statistically isotropic two-phase composites ($\mathbf{K}^r = \kappa^r\mathbf{I}$ where $\mathbf{I}$ denotes the second

 ISBN: 978-84-616-9216-3

order identity tensor), the Hashin-Shtrikman bounds for the effective thermal conductivity are given by

$$(\mathbf{K}^{HS})^{-} = \frac{\mathbf{K}_0\mathbf{K}_1 + 2\mathbf{K}_0(\mathbf{K}_0\phi_1 + \mathbf{K}_1\phi_0)}{2\mathbf{K}_0 + \mathbf{K}_0\phi_0 + \mathbf{K}_1\phi_1}, \quad (\mathbf{K}^{HS})^{+} = \frac{\mathbf{K}_0\mathbf{K}_1 + 2\mathbf{K}_1(\mathbf{K}_0\phi_1 + \mathbf{K}_1\phi_0)}{2\mathbf{K}_1 + \mathbf{K}_1\phi_1 + \mathbf{K}_0\phi_0},$$

Derivations of the Hashin-Shtrikman bounds have been improved and revised by many authors since they were originally devised ([10], [11]). In particular, Ponte-Castañeda & Willis ([6]), introducing a comparison material and under additional microstructure information represented by a two-point correlation function, derived a more general expression for n types inclusion phases that could be select independently of their spatial distribution.

However, in general these bounds use to appear in the literature to be merely stated (not derived) and it is often unclear how to construct such bounds when the material is not of simple type (e.g. isotropic spheres inside an isotropic host phase). For this reason, our aim in this work is to derive a direct way of constructing the Hashin-Shtrikman bounds for transversely isotropic composites from first principles in the thermal conductive setting. That is, given the volume fractions, thermal conductivity properties, spheroidal shapes of phases of the composite and their spatial distribution, we construct a procedure by which the Hashin-Shtrikman bounds could be obtained in a straightforward manner using the correct tensor basis set and the appropriate expressions for the Hill tensors. In particular in this respect, assuming homogeneous temperature conditions in the far field and by using the associated Green tensor, we exploit completely the uniformity of the Hill tensor, obtaining explicit expression derived for spheroidal inclusions and distributions instead of others given in integral form. We also exhibit the implementation of the constructions with several examples and compare them with the homogenized results obtained for a periodic composite.

Acknowledgements

The authors thank Prof. I.D. Abrahams (University of Manchester) for helpful discussions, Prof. P.A. Martin (Colorado School of Mines) for reading through an early draft and Prof. A.N. Norris (Rutgers University) for providing useful references on tensor bases.

This work has been partially supported by the project MTM 2011-24457 of the "Ministerio de Ciencia e Innovación" of Spain and the research group FQM-309 of the "Junta de Andalucía".

References

[1] Z. Hashin, S. Shtrikman, *A variational approach to the theory of the elastic behaviour of multiphase materials,* J. Mech. Phys. Solids **11** (1963) 127–140.

[2] K. A. Lurie, A. V. Cherkaev, *Exact estimates of conductivity of composites*, *Prc. Royal Soc. Edinburgh*, **A99**, (1984) 71–84.

[3] J. C. Maxwell, *Treatise on Electricity and Magnetism*, Clarendon, Oxford, 1973.

[4] G. A. Milton, *The Theory of Composites*, Cambridge University Press.

[5] F. Murat, L. Tartar, *H-Convergence*, in *Topics in the Mathematical Modelling of composite Materials*, ed. A. Cherkaev, R. Khon, Birkhäuser, Boston, 21–43.

[6] P. Ponte Castañeda, J.R, Willis, *The effect of spatial distribution on the effective behaviour of composite materials and cracked media, J. Mech. Phys. Solids* **43** (1995) 1919–1951.

[7] A. Reuss, *Calculation of the flow limits of mixed crystals on the basis of the plasticity of mono-crystals. Z. Angew. Math. Mech.* **9** (1929) 49–58.

[8] W. Voigt, *Ueber die Beziehung zwischeden beiden Elasticitätsconstanten, Annalen der Physik* **38** (1889) 573–587.

[9] L. Tartar, *Estimations fines de coefficients homogénéisés*, in *Ennio De Giorgi Colloquium*, ed. P. Kree, Research Notes in Mathematics 125, Pitman, London, 1985, 136–212.

[10] L.J. Walpole, *On bounds for the overall elastic moduli of inhomogeneous systems - I, J. Mech. Phys. Solids* **14** (1966) 151–162.

[11] J.R. Willis, *Bounds and self-consistent estimates for the overall moduli of anisotropic composites, J. Mech. Phys. Solids* **25** (1977) 185–202.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Bifurcations of the roots of a 6-degree symmetric polynomial coming from the fixed point operator of a class of iterative methods

Beatriz Campos[1], Alicia Cordero[2], Á. Alberto Magreñan[3], Juan R. Torregrosa[2] and Pura Vindel[1]

[1] *Instituto de Matemáticas y Aplicaciones de Castellón, Universitat Jaume I, Spain*

[2] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Spain*

[3] *Dpto. Matemáticas, Universidad Internacional de La Rioja, Spain*

emails: campos@uji.es, acordero@mat.upv.es, alberto.magrenan@gmail.com, jrtorre@mat.upv.es, vindel@uji.es

Abstract

We consider the 6-degree polynomial whose roots provide the fixed points of the operator associated to the (α, c)-family of iterative methods. We analyze the bifurcations of these roots in the (α, c)-plane and we show, in the bifurcation diagrams, which are the ranges of parameters α and c for which they are real or complex.

Key words: polynomial roots, bifurcation diagrams.

1 Introduction

Iterative methods are needed for solving most of the nonlinear equations because they are difficult or impossible to solve exactly by means of analytical methods. When they are applied on polynomials, they give rise to rational functions whose behaviors are not well known except in a narrow area. A possible way to extend these regions is by studying the dynamical behavior of the rational functions. In some previous papers, we focus on this second option and we have started with the dynamical study of Chebyshev-Halley family([2], [3], [4]), the King's family [1], the c-family [5] and, finally, the (α, c)-family which includes Chebyshev-Halley [7] and $c-$families.

The (α, c)-family is a two-parametric third-order class of iterative methods defined by:

$$z_{n+1} = z_n - \left(1 + \frac{1}{2}\frac{L_f(z_n)}{1-\alpha L_f(z_n)} + cL_f(z_n)^2\right)\frac{f(z_n)}{f'(z_n)}, \tag{1}$$

where

$$L_f(z) = \frac{f(z)\,f''(z)}{f'(z)^2}$$

and α and c are complex parameters. As we pointed before, this family includes Chebyshev-Halley family for $c = 0$ and c-family when $\alpha = 0$.

In the study that we are conducting, we note that the dynamical behavior of this family is much more complicated because it includes two parameters. We apply this family on quadratic polynomials $p(z) = z^2 + a$. For this polynomial, the operator $M_p(z, \alpha, c, a)$ associated to (1) is a rational function depending on three complex parameters: a, α and c.

Due to the Scaling Theorem is verified, we can obviate the parameter a and the operator $M_p(z, \alpha, c, a)$ is conjugated to:

$$O_p(z, \alpha, c) = z^3\frac{(1+z)^4(-2+2\alpha-z) + 4c(1+z(2-2\alpha+z))}{(1+z)^4(2\alpha z - 1 - 2z) + 4cz^3(1+z)^2 - 8\alpha c z^4}. \tag{2}$$

As we have said, iterative methods are used for finding roots of a nonlinear equation and, from a dynamical point of view, these roots are some of the fixed points of the operator associated with the method.

From this dynamical point of view, our main interest lies in finding the fixed points of the operator (2) and to study their behavior. For this operator we obtain the following fixed points: $0, \infty$ (that coincides with the roots of the polynomial after applying the Möebius map, see [6] for example), $z = 1$, that it is a strange fixed point, and six more strange fixed points that are the roots of a symmetric 6-degree polynomial.

In this paper we find analytically the exact roots of this 6-degree polynomial. We classify them in the $(\alpha, c)-$plane, dividing the plane in different regions depending on the number of real and complex roots they contain and we study the bifurcations of these roots when crossing the boundaries of different regions. We also show the bifurcation diagrams for different values of the parameter α, where real roots are depicted. Observe that, as we find regions of the plane (α, c) where all roots are real, we can determine real values of these parameters for which the iterative method converges to non desired real points that are not solutions of our problem.

2 Calculus of the fixed points

The fixed points satisfy $O_p(z, \alpha, c) = z$. The relation $O_p(z, \alpha, c) - z$ writes as:

$$O_p(z, \alpha, c) - z = -z(z-1)\frac{P(z, \alpha, c)}{(1+z)^4(2\alpha z - 1 - 2z) + 4cz^3(1+z)^2 - 8\alpha c z^4}$$

 ISBN: 978-84-616-9216-3

where $P(z,\alpha,c)$ is the 6-degree polynomial:

$$P(z,\alpha,c) = z^6+(7-2\alpha)z^5+(19-8\alpha+4c)z^4+(26-12\alpha+8c-8\alpha c)z^3+(19-8\alpha+4c)z^2+(7-2\alpha)z+1. \tag{3}$$

The roots of this polynomial are strange fixed points of the operator associated to the class of iterative methods. So, we are interested in finding these roots in the (α,c)-plane.

We can observe that $P(z,\alpha,c)$ is a symmetric polynomial; then, as $z=0$ is not a root, we can apply the change of variables

$$y = z + \frac{1}{z}, \tag{4}$$

that transforms the equation $P(z,\alpha,c)=0$ into the cubic one

$$y^3 + (7-2\alpha)y^2 + (16-8\alpha+4c)y + (12-8\alpha+8c-8\alpha c) = 0. \tag{5}$$

Following Cardano's method we first eliminate the square term by the substitution

$$y = w - \frac{(7-2\alpha)}{3}, \tag{6}$$

obtaining the reduced form of the equation:

$$w^3 + pw + q = 0, \tag{7}$$

where

$$p = \frac{1}{3}(-(-1+2\alpha)^2+12c), \quad q = \frac{-2}{27}(-1+2\alpha)^3 + 18(1+4\alpha)c.$$

By applying the change $w=u+v$, we observe that u^3 and v^3 are the solutions of the quadratic equation

$$w^2 + qw - \frac{p^3}{27} = 0. \tag{8}$$

Then,

$$u^3, v^3 = \frac{-q \pm \sqrt{q^2+\frac{4p^3}{27}}}{2} \Rightarrow u, v = \sqrt[3]{\frac{-q \pm \sqrt{\Delta}}{2}},$$

where

$$\Delta = q^2 + \frac{4p^3}{27} = \frac{16}{27}c\left(2\alpha(-1+2\alpha)^3 + c\left(-1+40\alpha+32\alpha^2+16c\right)\right).$$

The sign of Δ determines the character (real or complex) of the roots.

$\diamond$ If $\Delta > 0$, equation (7) has one real root and a pair of complex conjugate roots. The real root w_0 is:

$$w_0 = u + v = \frac{1}{3}\sqrt[3]{f(a,c)} + \frac{1}{3}\sqrt[3]{g(a,c)},$$

 ISBN: 978-84-616-9216-3

with

$$\begin{aligned} f(\alpha,c) &= (-1+2\alpha)^3+18(1+4\alpha)c+6\sqrt{3}\sqrt{c\left(2\alpha(-1+2\alpha)^3+c(-1+40\alpha+32\alpha^2+16c)\right)} \\ g(\alpha,c) &= (-1+2\alpha)^3+18(1+4\alpha)c-6\sqrt{3}\sqrt{c\left(2\alpha(-1+2\alpha)^3+c(-1+40\alpha+32\alpha^2+16c)\right)} \end{aligned}$$

and the corresponding real solution of (5) is:

$$y_0=\frac{(2\alpha-7)}{3}+\frac{1}{3}\sqrt[3]{f(\alpha,c)}+\frac{1}{3}\sqrt[3]{g(\alpha,c)}.$$

The expressions for the complex roots are:

$$w_1=\frac{-1}{6}(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)})+i\frac{\sqrt{3}}{6}(\sqrt[3]{f(\alpha,c)}-\sqrt[3]{g(\alpha,c)}),$$

$$w_2=\frac{-1}{6}(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)})-i\frac{\sqrt{3}}{6}(\sqrt[3]{f(\alpha,c)}-\sqrt[3]{g(\alpha,c)})$$

and the corresponding complex solutions of (5) are:

$$y_1=\frac{(2\alpha-7)}{3}-\frac{1}{6}(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)})+i\frac{\sqrt{3}}{6}(\sqrt[3]{f(\alpha,c)}-\sqrt[3]{g(\alpha,c)})$$

and

$$y_2=\frac{(2\alpha-7)}{3}-\frac{1}{6}(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)})-i\frac{\sqrt{3}}{6}(\sqrt[3]{f(\alpha,c)}-\sqrt[3]{g(\alpha,c)}).$$

$\diamond$ If $\Delta=0$, equation (7) has two real roots, one of them double. Their expressions are:

$$\begin{aligned} z_0 &= 2\sqrt[3]{-\frac{q}{2}}=\frac{3q}{p} \\ z_1 &= z_2=2-\sqrt[3]{-\frac{q}{2}}=-\frac{3q}{2p}. \end{aligned}$$

If $q=p=0$, then $z=0$ is a triple root. In this case, the corresponding solutions of (5) are:

$$y_0=\frac{(2\alpha-7)}{3}+\frac{3q}{p}$$

and

$$y_1=y_2=\frac{(2\alpha-7)}{3}-\frac{3q}{2p}.$$

$\diamond$ If $\Delta<0$, equation (7) has three real roots whose real expressions are, in a trigonometric form:

$$w_k=2\sqrt{-\frac{p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{-q}{2}\sqrt{\frac{-27}{p^3}}\right)+\frac{2k\pi}{3}\right),\quad k=0,1,2.$$

 ISBN: 978-84-616-9216-3

Figure 1: Curves C_- and C_+.

Let us notice that the curve C defined by

$$2\alpha\left(-1+2\alpha\right)^3+c\left(-1+40\alpha+32\alpha^2+16c\right)=0,$$

corresponding to $\Delta=0$, divides the $(\alpha,c)-$plane into regions corresponding to $\Delta>0$ and $\Delta<0$. We denote by C_- and C_+ the lower and upper branches of the curve C, respectively (see Figure 1). Summarizing, the solutions of equation (5) are:

$$\begin{aligned}
y_1(\alpha,c) &= \frac{1}{3}(2\alpha-7)+\frac{1}{3}\left(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)}\right)\\
y_2(\alpha,c) &= \frac{1}{3}(2\alpha-7)-\frac{1}{6}\left(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)}\right)-i\frac{\sqrt{3}}{6}\left(\sqrt[3]{f(\alpha,c)}-\sqrt[3]{g(\alpha,c)}\right)\\
y_3(\alpha,c) &= \frac{1}{3}(2\alpha-7)-\frac{1}{6}\left(\sqrt[3]{f(\alpha,c)}+\sqrt[3]{g(\alpha,c)}\right)+i\frac{\sqrt{3}}{6}\left(\sqrt[3]{f(\alpha,c)}-\sqrt[3]{g(\alpha,c)}\right).
\end{aligned}$$

Undoing the change of variables (4), the six roots of the polynomial (3) are obtained by solving the quadratic equation $z^2-z,\,y+1=0$, whose solutions are

$$z_\pm=\frac{y\pm\sqrt{y^2-4}}{2}. \tag{9}$$

That is, the six roots of (3) obtained by substituting y_1, y_2 and y_3 in (9) are:

$$\begin{aligned}
z_1(\alpha,c) &= \frac{y_1(\alpha,c)+\sqrt{y_1(\alpha,c)^2-4}}{2}, \qquad z_2(\alpha,c)=\frac{y_1(\alpha,c)-\sqrt{y_1(\alpha,c)^2-4}}{2},\\
z_3(\alpha,c) &= \frac{y_2(\alpha,c)+\sqrt{y_2(\alpha,c)^2-4}}{2}, \qquad z_4(\alpha,c)=\frac{y_2(\alpha,c)-\sqrt{y_2(\alpha,c)^2-4}}{2},\\
z_5(\alpha,c) &= \frac{y_3(\alpha,c)+\sqrt{y_3(\alpha,c)^2-4}}{2}, \qquad z_6(\alpha,c)=\frac{y_3(\alpha,c)-\sqrt{y_3(\alpha,c)^2-4}}{2}.
\end{aligned} \tag{10}$$

 ISBN: 978-84-616-9216-3

3 Bifurcations of the fixed points

In this section, we realize a complete study of the evolution of the six roots in the $(\alpha, c)-$plane, that is, we analyze how they bifurcate as parameters α and c vary.

Above results have divided the $(\alpha, c)-$plane into two regions separated by the curves C_- , C_+ and the axis $c = 0$. In the orange region (Figure 1) as $\Delta < 0$, equation (7) has three real roots and in the blue region, as $\Delta > 0$, equation (7) has one real root and a pair of complex ones. Then, these curves are bifurcation curves. But we have also to undo the change (9) in order to obtain the six solutions of (3); then, the real roots can originate a pair of complex roots if $y^2 - 4 < 0$. Therefore, we make $y = 2$ and $y = -2$ in order to find all bifurcation curves.

Let us consider

$$y_1(\alpha, c) = 2 \Rightarrow \frac{1}{3}(-7 + 2\alpha) + \frac{1}{3}\left(\sqrt[3]{f(\alpha, c)} + \sqrt[3]{g(\alpha, c)}\right) = 2.$$

Then, the hyperbola

$$C(\alpha) = \frac{2(-5 + 2\alpha)}{2 - \alpha}$$

is obtained. We consider the two branches of this curve and we denote by C_1 the branch for $\alpha < 2$ and C_2 the branch for $\alpha > 2$.

Similarly, making $y_1(\alpha, c) = -2$ we obtain $216\alpha c = 0$ and the curves $c = 0$ and $\alpha = 0$ are also bifurcation curves. The same separating curves appear by considering $y_2(\alpha, c) = \pm 2$ and $y_3(\alpha, c) = \pm 2$.

In Figure 2 the different regions separated by the bifurcation curves are shown. The fixed points can change from complex to real, or vice versa, when they cross these curves. We also show the number of real or complex roots in each region.

Now, let us analyze these bifurcations. We consider different fixed values for the parameter α and we vary the value of the parameter c, so that all regions are covered. In the bifurcation diagrams the roots z_1, z_2, z_3, z_4, z_5 and z_6 defined in (10) are depicted in red, yellow, magenta, orange, green and blue, respectively (Figures 3 to 7).

i) $\alpha = -1$. There are 6 real roots for negative values of c below the curve C_1. For $c = -14/3$ this curve is crossed; z_5 and z_6 reach the value 1 and become a pair of complex conjugate roots; after the bifurcation there are 4 real roots and two complex roots. When the value of c arrives to 0, another bifurcation occurs: the real roots z_3 and z_4 reach the value -1 and become a pair of complex conjugate roots, the pair of complex roots z_5 and z_6 take the value -1 and continue as a pair of complex conjugate roots. Then, at the bifurcation point we have the roots $z_1 = -0.208712$, $z_2 = -4.79129$, $z_3 = z_4 = z_5 = z_6 = -1$ and after the bifurcation there are two real roots (z_1 and z_2) and four complex roots. The bifurcation diagram is shown in Figure 3a.

ii) $\alpha = -0.01$. We consider this value for α in order to cover the little red region near the origin in Figure 2. The bifurcation diagram is the same as in the case above up to the

 ISBN: 978-84-616-9216-3

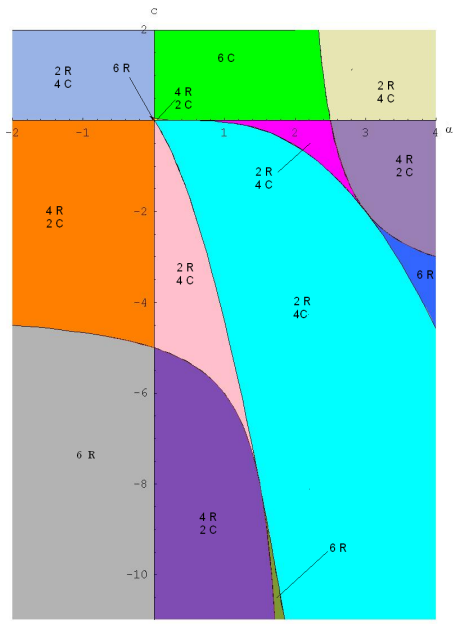
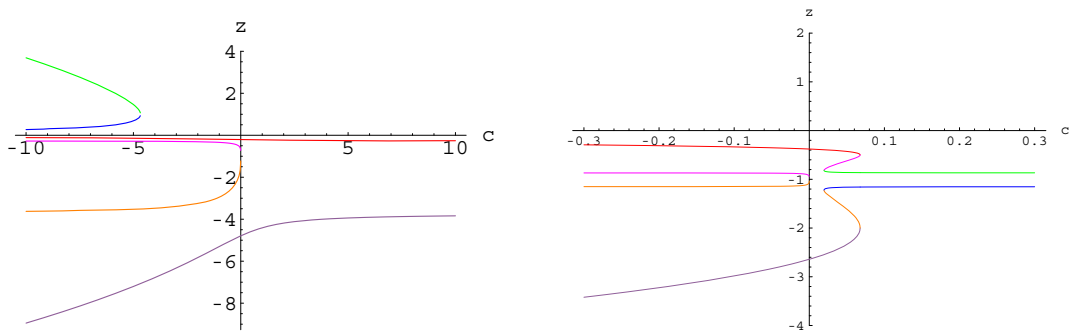


Figure 2: Zones in the plane (α, c) defined by the curves C_1, C_-, C_+, C_2, and the coordinate axes.

curve C_-. Then, we have 6 real roots for negative values of c below the curve C_2 and 4 real roots and two complex roots after crossing it. When the value of c arrives to 0, another bifurcation occurs and after the bifurcation there are 2 real roots (z_1 and z_2) and 4 complex roots. In this case, the curve C_- is reached for $c = 0.0195915$. Now, we enter the red region where there are 6 real roots and, increasing the value of c, we reach the curve C_+ for the value $c = 0.0677085$. After this bifurcation, z_1 and z_3 become a pair of complex roots and z_2 and z_4 become another pair of complex roots; z_5 and z_6 remain real roots. So, there a are 4 complex and 2 real roots after the bifurcation. The bifurcation diagram for negative values of c is similar to the previous case; we show in Figure 3b the detail of the bifurcation diagram for $-0.3 \leq c \leq 0.3$.

iii) $\alpha = 0.01$. The bifurcation diagram is shown in Figure 4a. For negative values of c below the curve C_1, we have 4 real roots and a pair of complex roots. Increasing the value of c we reach the curve C_1 finding a bifurcation: on the curve the roots are $z_1 = -0.145898$, $z_2 = -6.85412$, $z_{3,4} = -0.98999 \pm 0.141138i$, $z_5 = z_6 = 1$. After the bifurcation z_5 and z_6 become a pair of complex roots; so, there are two real and four complex roots.

We find a second bifurcation when we reach the curve C_- for the value $c = -0.0203924$. At this bifurcation point both pairs of complex conjugated roots become a double pair of complex numbers and after the bifurcation we continue having 2 real roots and 4 complex



(a) $\alpha = -1$. (b) $\alpha = -0.01$ and $-0.3 \leq c \leq 0.3$.

Figure 3: Bifurcations diagrams for $\alpha < 0$

roots.

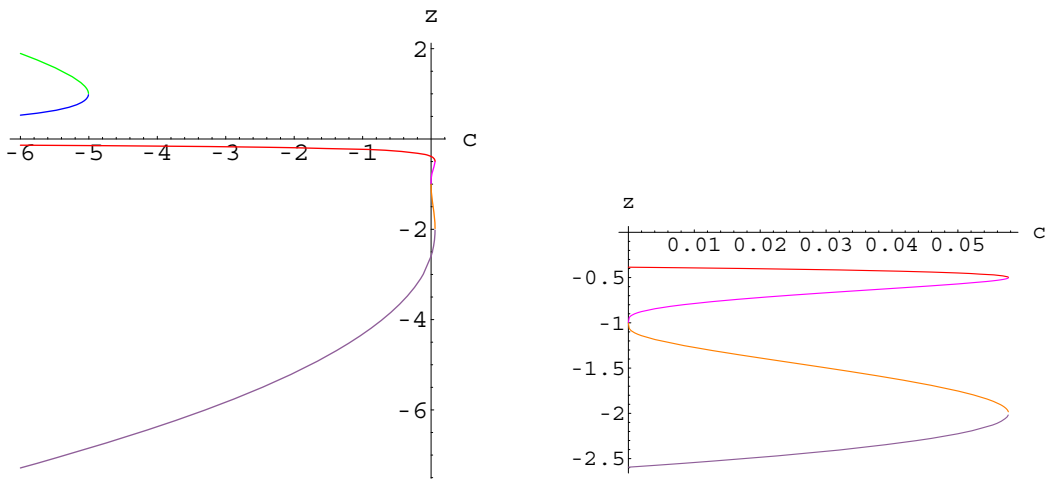
A third bifurcation occurs when crossing the curve $c = 0$. On the α axis we have $z_1 = -0.385419$, $z_2 = -2.59458$, $z_3 = z_4 = z_5 = z_6 = -1$. The four complex roots have become -1 (their real part become -1 and the imaginary part become 0). After the bifurcation z_3 and z_4 continue real while z_5 and z_6 become complex roots (see the detail of this bifurcation in Figure 4b). After this bifurcation we have 4 real and 2 complex roots corresponding to the yellow zone.

The last bifurcation we find in this direction is produced for $c = 0.0576924$ when the curve C_+ is reached. The roots z_2 and z_4 converge to the value -2 and become a pair of complex conjugated roots. Similarly, the roots z_1 and z_3 converge to the value -0.5 and become a pair of complex conjugated roots. Then, on the curve we have a pair of complex roots and two pairs of real double roots and after the bifurcation there are 6 complex roots.

iv) $\alpha = 0.6$. The difference of moving c in this case from the case before occurs when crossing the α axis. We have 2 real and 4 complex roots for $c < 0$ and we have 6 complex roots for $c > 0$. For $c = 0$ the roots are -1 with multiplicity 4 and the pair of complex $z_{5,6} = \frac{1}{10}(-9 \pm \sqrt{19}i)$. At this bifurcation, the real roots z_1 and z_2 reach the value -1 and become a pair of complex roots; the complex roots z_3 and z_4 take the value -1 but after the bifurcation they continue being complex; finally, the complex roots z_5 and z_6 continue being complex. Therefore, for $c > 0$, there are 6 complex roots. The bifurcation diagram is shown in Figure 5.

v) $\alpha = 1$. Starting with negative values of c the two first bifurcations are similar to the previous case. Now, we continue increasing the valor of c up to value $c = 0$. For this value of c we have two complex roots and a real root -1 with multiplicity 4. There are a pair of complex roots whose imaginary part becomes 0 and the real part becomes -1 but after the

 ISBN: 978-84-616-9216-3



(a) $\alpha = 0.01$. (b) $\alpha = 0.01$ and $0 \leq c \leq 0.06$.

Figure 4: Bifurcation diagrams for $\alpha = 0.01$.

bifurcation they remain complex and the two real roots that reach the value -1 become a pair of complex roots. Therefore, after this bifurcation, for $c > 0$, there are six complex roots. As we only depict the real roots, the bifurcation diagram is similar to the diagram of Figure 5.

vi) $\alpha = 1.5$. At the point $(1.5, -8)$ the curves C_1 and C_- intersect. There are four real and two complex roots below this point.

At the bifurcation point the real roots z_3 and z_4 reach the value 1 and become a pair of complex roots and the complex roots z_5 and z_6 also take the value 1 but they continue complex after the bifurcation. The rest of behavior as c increases is the same as the case above and the bifurcation diagram is also similar to that of Figure 5 and we do not show it.

vii) $\alpha = 1.9$. We consider a value lower than 2 in order to cross the curve C_1 from the violet to the dark green zone, being that $\alpha = 2$ is one asymptote of the curve C_1. In the violet region z_3 and z_4 are complex and the rest of the roots are real.

Increasing the value of c the curve C_1 is reached for $c = -24$. At this bifurcation point the pair of complex z_3 and z_4 become real and after the bifurcation there are 6 real roots.

Now, we reach the curve C_- for $c = -11.45225$; at the bifurcation point there are two double real roots that become pairs of complex conjugate roots after the bifurcation; then, after crossing C_- there are 2 real and 4 complex roots. Now, as we are to the left of the asymptote $c = 2$, increasing the value of c from this point we find the same bifurcations as the case before. The bifurcation diagram is shown in Figure 6a.

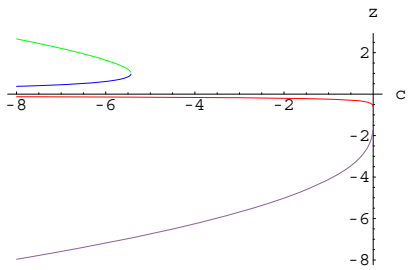


Figure 5: $\alpha = 0.6$

viii) $\alpha = 2.25$. The bifurcations for negatives values of c are as in the case for $\alpha = 1.9$. But now, once the α axis is crossed, where there are 6 complex roots, we reach the curve C_2 for $c = 4$. At the bifurcation point the roots $z_1 = z_2 = 1$ and after the bifurcation there are 2 real and 4 complex roots. The bifurcation diagram is shown in Figure 6b.

ix) $\alpha = 3$. For negative values of c we have the same bifurcation as the case before. For the $c = -2$ we pass directly to the region with 4 real and 2 complex roots. When arriving to the value $c = 0$, we find another bifurcation: the roots z_1 and z_2 have reached the value -1 and the complex z_3 and z_4 take also the value -1. After the bifurcation, the roots z_1 and z_2 become a pair of complex conjugate roots and z_3 and z_4 continue being complex; then, there are 2 real roots and 4 complex roots. The bifurcation diagram is shown in Figure 7a.

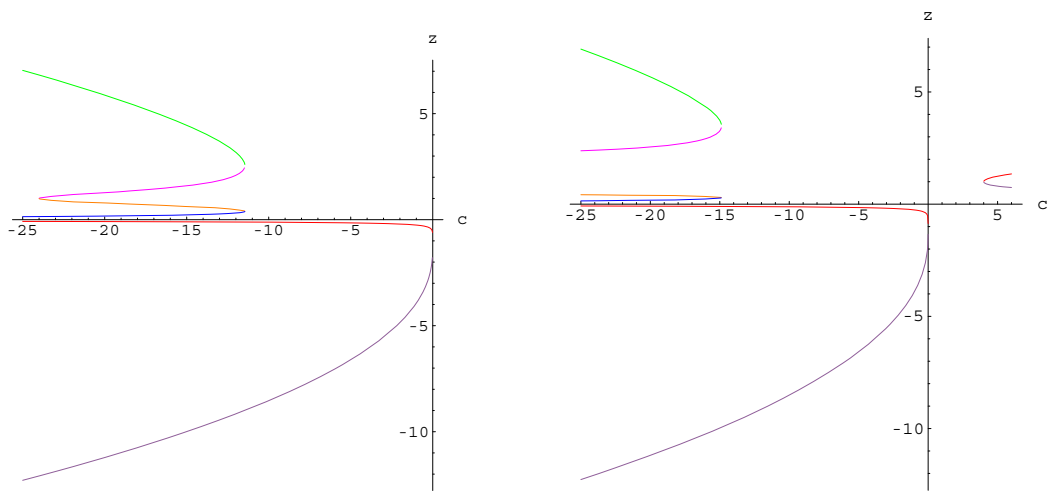
x) $\alpha = 4$. We find the same bifurcation as the case before when crossing from the dark green zone to the cyan zone for $c = -37.3452$. In this case, increasing the value of c we cross the curve C_+ for the value $c = -4.59229$. After this bifurcation we have 6 real roots. For the value $c = -3$ the curve C_2 is crossed. After this bifurcation, the real roots z_3 and z_4 become a pair of complex roots. So, we have 2 real and 4 complex roots.

For $c = 0$, we have another bifurcation: z_1 and z_2 take the value -1 and become a pair of complex roots. The complex z_3 and z_4 take the value -1 at the bifurcation point but they continue being complex for positive values of c. Therefore, for $c > 0$ there are 2 real and 4 complex roots. In Figure 7b we show the part of the bifurcation diagram for $-6 \leq c \leq 2$. For values around $c = -37.3452$ we have a similar diagram as in the Figure 7a for values around $c = -23.4375$.

4 Final remarks

If we look carefully the above diagrams we can observe the bifurcations for different values of c. For example, for positive values of c two behaviors are distinguished:

 ISBN: 978-84-616-9216-3



(a) $\alpha = 1.9$. (b) $\alpha = 2.25$.

Figure 6: Bifurcations diagrams for α around 2.

- For $c > \frac{27}{256}$ there are two real and four complex roots if $\alpha < 0$ or α is at the right of the curve C_2.
- If $0 < c < \frac{27}{256}$ there are two real roots for $\alpha < 0$ at the left of the curve C_-, six real roots when $\alpha < 0$ at the right of the curve C_-, four of which remain for $\alpha > 0$. All the roots become complex when α is between the curves C_+ and C_2 and two of them become real for α at the right of the curve C_2.

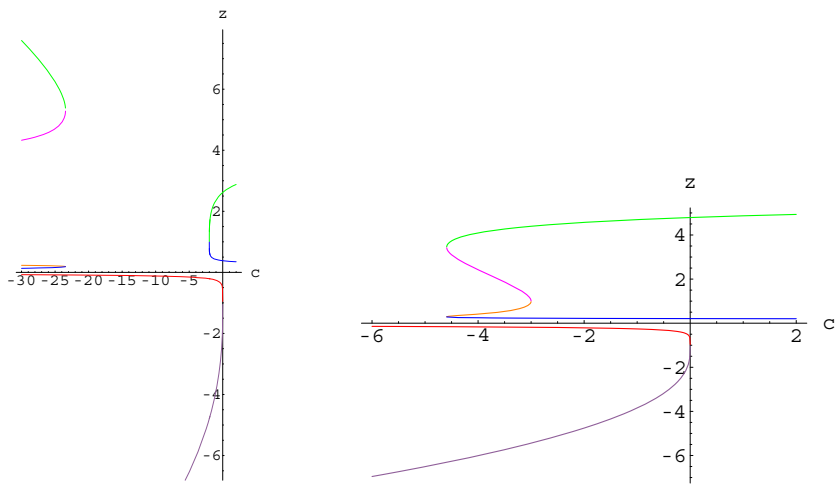
A similar analysis can be obtained for negative values of c. This analysis will help us to select the members of the class of (α, c) iterative methods to find the real roots of a nonlinear equation, in terms of their stability and reliability.

Acknowledgements

This work has been supported by Ministerio de Ciencia y Tecnología MTM2011-28636-C02-02.

References

[1] A. Cordero, J. García-Maimó, J.R. Torregrosa, M.P. Vassileva and P. Vindel, *Chaos in King's iterative family*, Applied Mathematics Letters 26 (2013), pp. 842—848.



(a) $\alpha = 3$. (b) $\alpha = 4$

Figure 7: Bifurcations diagrams for $\alpha \geq 3$

[2] A. Cordero, J.R. Torregrosa and P. Vindel. *Dynamics of a family of Chebyshev-Halley type methods*. Applied Mathematics and Computation 219 (2013), pp. 8568—8583.

[3] A. Cordero, J.R. Torregrosa and P. Vindel, *Period-doubling bifurcations in the family of Chebyshev—Halley type methods*, International Journal of Computer Mathematics 90(10) (2013) pp. 2061–2071.

[4] A. Cordero, J.R. Torregrosa and P. Vindel. *Bulbs of period two in family of Chebyshev-Halley type methods on quadratic polynomials*. Abstract and Applied Analysis Volume 2013, ID 536910 10 pages.

[5] B. Campos, A. Cordero, J.R. Torregrosa and P. Vindel. *Dynamics of the family of c-iterative methods*, International Journal of Computer Mathematics, 2014. doi:10.1080/00207160.2014.893608

[6] A. Douady and J.H.Hubbard, *On the dynamics of polynomials-like mappings*, Ann. Sci. Ec. Norm. Sup. (Paris) 18 (1985), pp. 287–343.

[7] N. Osada, *Chebyshev-Halley methods for analytic functions*, Journal of Computational and Applied Mathematics 216(2) (2008), pp. 585–599.

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Optimizing the Performance of Financial Applications on Heterogeneous Architectures

Emilio Castillo[1], Cristobal Camarero[1], Ana Borrego[2] and Jose L. Bosque[1]

[1] *Department of Ingeniería Informática y Electrónica, Universidad de Cantabria*

[2] *Technology Division, Grupo Santander - Produban*

emails: emilio.castillo@unican.es, cristobal.camarero@unican.es, aiborrego@produban.com, joseluis.bosque@unican.es

Abstract

The use of high performance computing systems to help making the right investment decisions in financial markets is an open research field where multiple efforts have being carried out during the last few years. Specifically, the HJM model has a number of features that make it well suited for implementation on massively parallel architectures. This paper presents a MultiCPU and MultiGPU implementation of the HJM model that improve both the response time and throughput. The experimental results reveal that the proposed architectures achieve good speed up and scalability, and optimize the power consumption and cost/performance ratio.

Key words: Heterogeneous Computing, MultiGPU, financial applications.

1 Introduction

Simulation is getting increasingly important in financial markets as one of the best techniques for improving the accuracy of investments. Also in this environment, the shorter the time in getting to an accurate assessment, the better. A microsecond or nanosecond faster than the rest of the players can lead to create market instead of just being on the market.

Financial simulations based on Monte Carlo methods have been used for many years thanks to their intrinsic parallelism. A Monte Carlo method is an algorithm that solves a problem through the use of statistical sampling to obtain numerical results [11]; typically it is necessary to run simulations many times over in order to obtain the distribution of an unknown probabilistic entity.

Monte Carlo methods have a number of properties that make then especially suitable for implementation on massively parallel environments [9, 2]. These include the data independence, that enable domain-based parallelization, with a high degree of parallelism, ie can generate a large number of fine-grained tasks or a few coarse-grain tasks. This property greatly favours the application scalability, while allowing an adequate distribution of the workload in both homogeneous and heterogeneous environments, which has a large impact on the final performance. Also the overhead due to synchronization or communication between processes or threads is minimal.

Specifically this work address the optimization of financial applications that allow a prediction of risk over time, for financial derivative products, particularly in multi-value environments. The selected model is the Heath Jarrow Morton (HJM) framework [7, 8]. Therefore, this paper presents a new and efficient implementation of the HJM Model, which has a high computational cost, in highly scalable, heterogeneous and cross-platform environments. In particular optimization techniques and code parallelization used in homogeneous environments (multicore architectures) as well as in heterogeneous environments (GPUs).

The HJM Model used in this paper is based on Monte Carlo methods. These methods have been widely used over time in many different fields, financial, engineering and scientific. For instance, [5] presents an implementation of a Monte Carlo model to estimate the current value of an European option for future purchase in the financial derivatives market, based on the Black-Scholes model. The implementation was done in four very different computer systems: A multicore with shared memory, a cluster with MPI, a CUDA program running on a GPU and a cluster of FPGAs where the most time consumed computations were implemented in VHDL. Similarly [12] presents the design and implementation of a parallel version of a Monte Carlo method in a FPGA-based supercomputer, called Maxwell, of Edinburgh University [3]. The FPGA-based implementation is compared with other environments with various GPUs and conventional processors.

On the other hand, [1] also uses clusters of CPUs and GPUs to implement the calculation of the price of European options. They compare different systems and implementations in terms of performance and power consumption. Many financial applications rely on solving systems of sparse linear equations. As for example, [6] proposes the design of a number of iterative methods for solving equations, based on the Krylov subspace on GPU architectures. In this work, the proposed approaches are validated by solving the partial differential equations of the Black-Scholes model.

As far as we know this is the first paper where the HJM model is implemented on a massively parallel architecture, like the proposed in this paper. Additionally, this paper proposes a study of the performance of this kind of applications in heterogeneous environments, from two different points of view: the improvement of performance (both response time and throughput) and scalability, as both are important in financial applications. Finally, a study on the power consumption and cost of these architectures is also shown.

 ISBN: 978-84-616-9216-3

2 Interest Rate Models

During the past three decades, derivatives have become increasingly important in the world of finance. A derivative can be defined as a financial instrument whose value depends on the values of other, more basic underlying variables. Very often the variables underlying derivatives are the prices of traded assets. Some major developments have occurred in the theoretical understanding of how derivative asset prices are determined, and how these prices change over time, led to the use of advanced mathematical methods. Models and numerical procedures based on the original Black-Scholes assumptions [4] are straightforward. However they have simplistic approaches and assumptions when tackling exotic options.

Therefore a number of alternative new models have since been introduced to attempt to solve this problematic. These models, such as the Hull White, the Vasicek, the Cox Ingersoll and Ross model, incorporate a description of how interest rates change through time. For this reason, they involve the building of a term structure, typically based on the short term interest rate r_t. The main advantage of these methods lies in the possibility of specifying r_t as a solution to a Stochastic Differential Equation. This allows, through Markov theory, to work with the associated Partial Differential Equation and to subsequently derive a rather simple formula for bond prices. This makes them widely suited for valuing instruments such as caps, European bond options and European swap options.

However, they have some limitations and all lead to the same drawback when solving interest rate products: the fact that they use only one explanatory variable (r_t) to construct a model for the entire market. It proves insufficient to realistically model the market curve, which appears to be dependent on all the rates and their different time intervals. Consequently, these models cannot be used for valuing interest rate derivatives such as American-style swap options and structures notes, as they introduce arbitrage possibilities.

2.1 Heath-Jarrow-Morton (HJM) framework

The most straightforward solution to the above problem should include the use of more explanatory variables: long and medium term rates. The Heath Jarrow Morton framework uses one representative short term rate, a middle term rate, and finally a long term interest rate [7, 8]. It chooses to include the entire forward rate curve as a theoretically infinite dimensional state variable. Unlike other models, this model can match the volatility structure observed today in the market, as well as at all future times.

The Heath-Jarrow-Morton framework is a general framework to model the evolution of interest rates. It describes the behaviour of the future price (in time) of a zero coupon bond $B(t,T)$ paying 1 unit of currency at time T, and it provides a consistent framework for the pricing of interest rate derivatives. The model is directly calibrated to the currently observed yield curve, and is complete in the sense that it does not involve the market price of interest rate risk.

The key aspect of HJM lies in the recognition that the drifts of the no-arbitrage evolution of certain variables can be expressed as functions of their volatilities and the correlations among themselves, so no drift estimation is needed. HJM-type models capture the full dynamics of the entire forward rate curve. In practice however, we will not work with a complete, absolutely continuous discount curve $B(t,T)$, but will instead construct our curve based on discrete market quotes, and will then extrapolate the data to make it continuous. Given the zero-coupon curve $B(t,T)$, there exists a forward rate $F(t,u)$ such that:

$$dF(t,T) = \mu(t,T)dt + \sigma(t,T)dWt_t^P \tag{1}$$

The HJM model has the serious disadvantage that it cannot be represented as recombining trees. In practice, this means that it must be implemented using Monte Carlo Simulations. Therefore, it has a very high computation time so it is important to use high performance architectures to minimize response times.

3 Graphics Processing Unit (GPU)

The GPU used in this work is a NVIDIA Tesla Kepler K20, with GK110 microarchitecture. The goal of the Kepler architecture focuses, not only on performance, but also on efficiency and programmability. It comprised 7.1 billion transistors, with 13 SMX Steaming Multiprocessor which contains 192 CUDA cores each one, so it has 2496 CUDA cores. It has a peak performance of 1.17 and 3.52 TFlops on double and single precision operations respectively. It also has 5 GBytes of GDDR5 memory, with a bandwidth of 208 GBytes per second. The architecture presents two new important features:

- **Dynamic Parallelism**, enables the Kepler GK110 GPU to dynamically spawn new threads by adapting to the data without going back to the host CPU. This effectively allows more of a program to be run directly on the GPU, as kernels now have the ability to independently launch additional workloads as needed. Any kernel can launch another kernel and can create the necessary streams, events, and dependencies needed to process additional work without the need for host CPU interaction.

- **Hyper-Q**, enables multiple CPU cores to launch work on a single GPU simultaneously. This feature allows 32 simultaneous hardware managed connections between the host and the GPU. Hyper-Q allows connections for CUDA streams, MPI processes and even threads from within a process. Legacy MPI applications were created to run on multi-core, and thereby, the amount of work in each MPI process is insufficient to fully occupy the GPU. One solution is to issue multiple MPI processes to concurrently run on the GPU, but it can produce false dependencies among them. Hyper-Q removes false dependency bottlenecks and increases speed at which MPI processes can be moved from the host to the GPU.

 ISBN: 978-84-616-9216-3

4 Optimization of the HJM Model

4.1 Analysis and Optimization of Sequential Code

The starting point is a sequential code that implements a multidivise prediction risk values model based on HJM [10], using a Monte Carlo method. This code was implemented in C++ language with the Intel MKL library. On this version a code profile using gprof and Valgrind has been done. The profile has been performed both with and without MKL to verify the impact of this library on performance. This profile shows that the 41.67% of the runtime is spent in a function of the MKL library which executes a exponential function. The remaining time is consumed mainly in other vector operations. Specifically the *operators*, a set of functions that performs simple operations on all the elements of several vectors that are calculated in a step of the simulation.

Another remarkable aspect is that the use of MKL library has a significant effect on performance, provided that use Intel processors. The execution time is reduced by 48%, reaching a speedup of 1.92 compared to the version without MKL. This improvement comes from both optimizations performed in the own library and the fact that it uses multi-threading, thus it is taking advantage of all processors in the system (two in this study).

Likewise Valgrind revealed a large number of memory conflicts due to the compiler uses memory areas are mapped to the same cache blocks. This produces *Cache Jamming*, consisting of two variables is constantly overwritten in the cache, resulting in a large number of replacements and thereby causing a strong performance degradation. Changing the memory allocation scheme of the variables involved this effect has been eliminated.

Finally, since the application uses several arrays with a large number of double-precision data (as many as the number of paths), the effect of the cache in the application performance has studied. Thus, the runtime has been measured on a processor with the same architecture and clock frequency, but a size of second-level cache (L2), which is 3 times higher per core. The results shows that the improvements, for commonly used sizes, are around 6%, in the response time.

4.2 Replacing MKL Library

The previous section showed that the use of the MKL library has a strong impact on application performance. However, it also has two major problems: the cost is very high and only takes advantage on Intel processors limiting code portability. Therefore, it is proposed to search an open source solution, to replace the functions of MKL used: exponential and division of floating point numbers in double precision. Alternatively, an approach based on SLEEF (SIMD Library for Evaluating Elementary Functions), an open source library, and AVX instructions is proposed. Two different approaches have been developed, the first based on a single thread and using vector instructions, and the second using multi-threading.

The version that uses only SLEEF with AVX instructions, down time with respect to the original version without MKL, but nonetheless takes around 20% more than using MKL. This is because SLEEF are using a single thread, while MKL adjusts the number of threads to data size. If a version SLEEF and multi-threading in the same areas of MKL is used, a very close result is obtained only 3% worse. MKL uses highly optimized routines with details of the processor architecture that are not public, so you get exactly the same performance is a non trivial task. This analysis shows an interesting conclusion: MKL can be replaced by an alternative open source without losing performance just allowing generate a more portable and less economic cost code.

4.3 CUDA Implementation

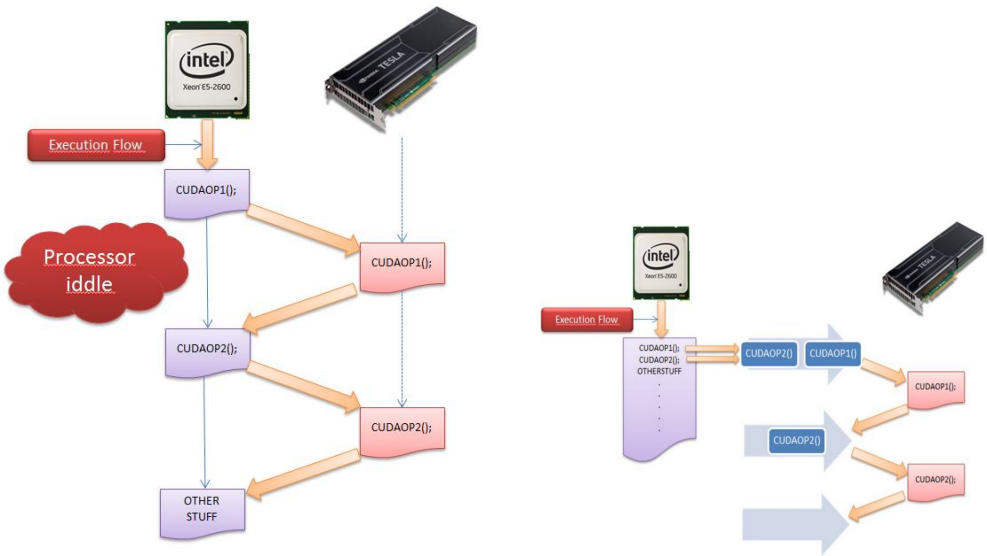
A first aspect to analyse is the communication between CPU and GPU as it is one of the main bottlenecks in the Host-Device programming model. The application is iterative, thus it performs a series of calls to CUDA kernels, one for each step of the simulation. The kernels execute the most most computational cost operations, such as exponential and division of floating point numbers in double precision, on large data vectors which are independent from each other.

A detailed analysis of the data dependencies between successive iterations shows that the results of the partial vector operations, performed at each simulation step, are not needed until the end of the execution, and hence they are always stored in the memory of the GPU. This has a double impact on performance: synchronization points between CPU-GPU are avoided and the transfer of information between the two devices is minimized. This is implemented through the use of *CUDA streams*. Each call to a CUDA operation is queued into a stream and the application can continue executing on the CPU asynchronously. The stream manages the execution of CUDA kernels while the CPU is computing the control structures and queues new CUDA kernels operations. The CPU waits for the GPU only when reading the final results, rather than once per transaction.

Figure 1(a) shows the execution flow of the synchronous case. When a CUDA operation is running on the GPU, the processor remains idle waiting until it ends. The processor idle time, could request new work to the GPU or perform independent CPU tasks which do not require pending results. On the other hand, 1(b) shows how the CPU queues a CUDA task in a special buffer, the *stream* and continues running other part of the code. The synchronization is only needed when a transfer of input data or results is essential.

Once the migration to he improvement obtained in response times was not as good as expected. This behaviour is explained because the operating system did not have the driver loaded in persistent mode, ie, the driver is only loaded into memory when a process needs to access the GPU. This causes the driver initialization time to accumulate the response time of the application. The solution was to load the driver in persistence mode to always remain in memory.

 ISBN: 978-84-616-9216-3

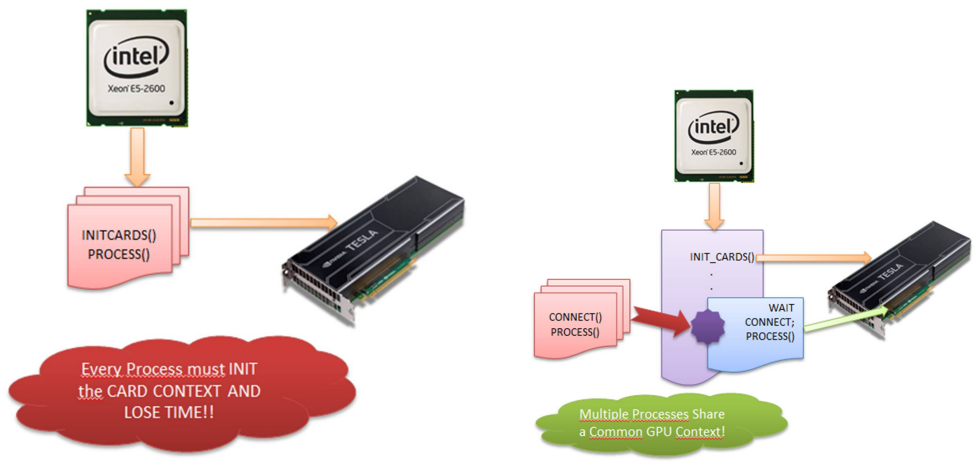


(a) Synchronous CPU-GPU Execution (b) Asynchronous CPU-GPU Execution

Figure 1: Synchronous vs. Asynchronous CPU-GPU execution with CUDA

Another overhead in the time is the CUDA initialization where the CUDA driver creates the memory maps, initializes registers and contexts, and finally loads the code into the GPU (Figure 2(a)). All of these steps (except the last, of course) may be made prior to the execution of the application. To solve this problem a *Client-Server* architecture has been designed, based on UNIX Domain Sockets (Figure 2(b)). The *Server* initializes the GPU and is listening on a socket, awaiting execution requests coming from the *Clients*. The *Clients* must know the specific port to communicate with the *Server*, and through this port sends the name of the file containing the kernel to be run. The *Server* runs the kernel on the GPU and returns the result to the *Client*. With this architecture, initialization is performed only once, at boot time of the machine. Thus the individual processes prevent overload time. The NVIDIA K20 GPU has an initialization time of about 100 ms. The execution times of kernels in this application are about 40 ms. It is therefore evident that the overhead introduced by the initialization has a strong impact on the application response time.

Finally, it is important to highlight that the implementation is Multi-CPU and Multi-GPU, ie, supports the execution of a single job on multiple GPUs in parallel. This can be specified as an input parameter to specify the maximum number of GPUs that can be used in each run. In the case of using more than one GPU workload is distributed statically, ie the workload is distributed at the beginning of the execution. Moreover, the distribution is homogeneous, i.e. the workload is evenly distributed among all the GPUs in the system.



(a) Without Server: one initialization per process

(b) With Server: one initialization at machine boot

Figure 2: Execution model with and without Server

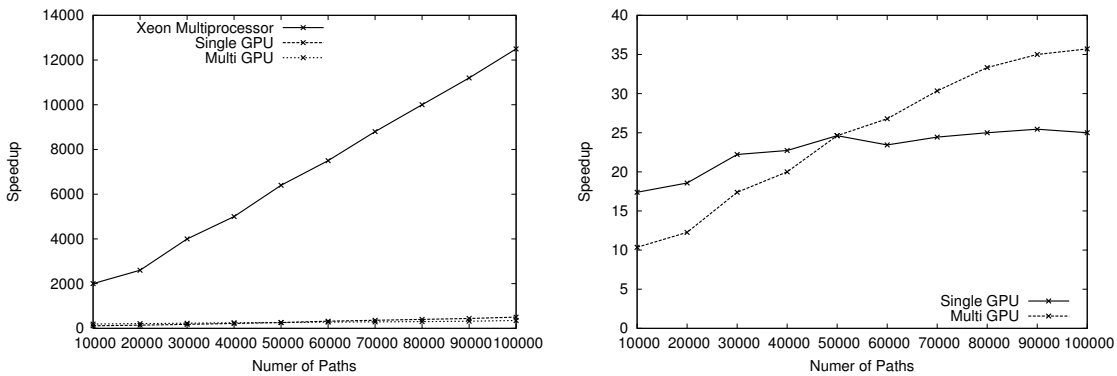
5 Experimental Results

This section presents a set of experimental results. The main objectives of these experiments are to perform an study of the performance of the proposed approach, varying some parameters of the test cases used, such as the number of paths executed; and to analyse in depth the performance of Multi-GPU environment, both in terms of response time of a single instance of the problem, and when the task throughput.

The experiments have been developed on a Intel server with a dual Intel Sandy Bridge E5-2620 processors with 6 cores each one, at 2 GHz. The server has hardware support for 24 threads, 15 MB of L3 cache memory and 16 GB of DDR3 main memory. The system runs a Ubuntu 10.04 Linux operating system, and has the CUDA 5 and Intel MKL Library. The server comprises two NVIDIA Kepler K20 GPUs with 2496 cores, 5 DDR5 GB of memory, with a peak performance of 3.52 TFlops on single precision and 1.17 TFlops on double precision operations. Each GPU has its own dedicated PCI-express 3.0 bus between the GPU and the CPU, to avoid collisions in the access to the bus.

All results presented in this section refer to the implementation in double precision, since in the initial requirements are considered the most interesting. The metric used in all cases is the *response time*, defined as the total execution time since the application is launched until results are obtained. Therefore includes both computing time, such as communication between CPU and GPU, for initialization, reading and writing operands and results. The times are always expressed in milliseconds to allow a better comparison. The results presented are always the average obtained from 10 independent runs.

 ISBN: 978-84-616-9216-3



(a) Response Times (b) Speedups

Figure 3: Speedups of Single and Multi GPU systems vs. Xeon processor

Table 1: Response Times and Speedups

Number of Paths	Xeon Processor	Single GPU	Multi GPU	Speedup Single GPU vs. Xeon	Speedup Multi GPU vs. Xeon	Speedup Multi GPU vs. SingleGPU
10000	2000	115	193	17.39	10.36	0.60
20000	2600	140	212	18.57	12.26	0.66
30000	4000	180	230	22.22	17.39	0.78
40000	5000	220	250	22.73	20.00	0.88
50000	6400	260	260	24.62	24.62	1.00
60000	7500	320	280	23.44	26.79	1.14
70000	8800	360	290	24.44	30.34	1.24
80000	10000	400	300	25.00	33.33	1.33
90000	11200	440	320	25.45	35.00	1.38
100000	12500	500	350	25.00	35.71	1.43

To evaluate the performance of the application, two kinds of experiments have been developed. The first experiment analyses the behaviour of the application's response time in a heterogeneous system. To do this, three different environments and implementations have been used:

- **Multi-Thread** application running on a multiprocessor with 12 cores.
- **Single GPU**, a heterogeneous Host+CUDA application running on a single GPU.
- **MultiGPU**, a heterogeneous Host+CUDA application, running on two GPUs.

In these scenarios, several tests were performed by changing the size of the problem, which is determined by the number of paths to evaluate. The metric used is the total response time of the application, including the communication time between CPU and GPU. Finally, the speedups of heterogeneous environments are computed. The results are presented in the table 1.

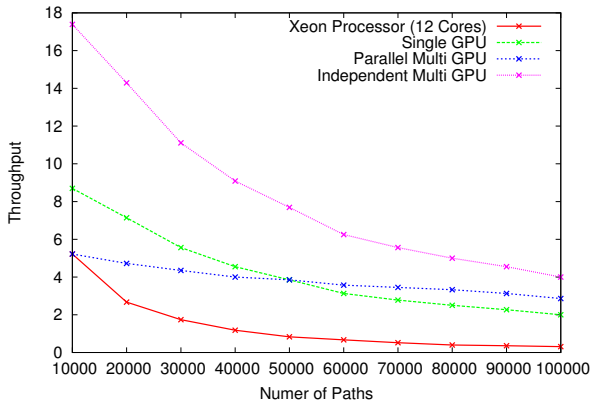


Figure 4: Throughput of different configurations

The first result that is important to highlight is the large reduction in response times that occurs when using the heterogeneous system, as can be seen in Figure 3(a). On the other hand, Figure 3(b), presents the behaviour of the speedups of the heterogeneous environments compared with the multiprocessor as problem size increases. In that figure, it can be seen that bellow 50.000 paths the speedup of the single GPU is significantly higher than MultiGPU. This behaviour is because the workload is too small and then the advantages of using two GPUs simultaneously can not improve the overhead to manage them. However, as the workload grows, the benefits of using two GPUs in parallel outperform this overhead, and this behaviour becomes more noticeable as the problem size grows.

The second experiment the metric used is not the response time of a single instance of the problem, but the throughput of the system, ie the number of tasks that can be completed in a certain time interval. Therefore multiple independent instances of the same problem are running simultaneously. In the case of Xeon multiprocessor these instances are not parallelized (ie, each job runs on a single core, so it can run 12 instances simultaneously). Figure 4 shows the throughput results obtained in all these available systems, as the number of paths increases. It can to be noticed that the environment with higher throughput is the Multi GPU, but using each GPU independently on a single instance of the problem. Furthermore, it is interesting to highlight that only one K20 GPU performs more operations per second than 12 Xeon cores running independent simulations.

Finally, the results achieved in the throughput with the GPUs, have a significant economic impact both the cost and power consumption. With the use of a single server with two E5-2620 processors and two GPU cards K20 performing simulations in parallel, it is possible to replace 10 servers without GPU. Therefore, it can be highlight that the heterogeneous architecture achieves a savings of 3.75 times in the power consumption as well as an initial investment in equipment 5.45 times lower.

 ISBN: 978-84-616-9216-3

6 Conclusions and Future Work

The most important and general conclusion to highlight is that the financial models based on Carlo methods, such as the HJM, have qualities that make them especially suitable for implementation on massively parallel architectures, especially in Multi-GPU platforms. Indeed, the massive data parallelism along with data independence allows to squeeze the full potential of the GPUs. Furthermore, these model minimize communication between CPU and GPU, that is one of the major bottle-necks in this architecture. Finally, this data independence also allows a balanced distribution of workload and offers excellent properties with regard to scalability.

This suitability is proven in the experimental results of response times and throughput presented in this paper. To summarize, it is noteworthy that a heterogeneous architecture with an NVIDIA Kepleer K20 GPU can achieve a speedup of more than 18 over the best version on CPU. Furthermore it has been shown that this architecture provides excellent scalability: the higher the workload, the better the speeup is, reaching up to 25 in the experiments presented in this paper. Finally, it is worth mentioning the low consumption of these architectures, as well as its excellent cost/performance ratio.

In Multi-GPU environments, the workload is the key parameter when deciding if the application runs on a single GPU or use several in parallel. The experimental results for the HJM model show that the use of the two GPUs in parallel is profitable from a workload paths 50,000. In more complex models with a higher cost of computation, this value can vary substantially.

Future work includes providing the MultiGPU environment with a load balancing mechanism that allows a heterogeneous distribution between GPUs with different performance. Likewise, other accelerator architectures such as Intel Xeon Phi will be explored.

Acknowledgements

The authors would like to express their gratitude to **François Friggit** of Banco Santander who inspired and motivated this challenge as a real business case and provided all necessary assistance to carry out this work.

The development of this paper has been partially supported by the Spanish Ministry of Education and Science, grant TIN2010-21291-C02-02, Consolider CSD2007-00050 and CAPAP-H network TIN2011-15734-E as well as by the HiPEAC European Network of Excellence.

References

[1] L.A. Abbas-Turki, S. Vialle, B. Lapeyre, and P. Mercier. High dimensional pricing of exotic european contracts on a gpu cluster, and comparison to a cpu cluster. In *Parallel Distributed Processing, IEEE International Symposium on (IPDPS)*, pages 1–8, 2009.

[2] Virat Agarwal, Lurng-Kuo Liu, and David A. Bader. Financial modeling on the cell broadband engine. In *22nd IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2008, Miami, Florida USA, April 14-18, 2008*, pages 1–12, 2008.

[3] Rob Baxter, Stephen Booth, Mark Bull, Geoff Cawood, James Perry, Mark Parsons, and Arthur Trew. Maxwell - 64 fpga supercomputer. *Engineering Letters*, 16(3), 2008. Special Issue: High Performance Reconfigurable Systems.

[4] Fishcer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637, 1973.

[5] Javier Castillo, José Luis Bosque, Emilio Castillo, Pablo Huerta, and José Ignacio Martínez. Hardware accelerated montecarlo financial simulation over low cost fpga cluster. In *IPDPS*, pages 1–8, 2009.

[6] A. Gaikwad and I.M. Toke. Parallel iterative linear solvers on gpu: A financial engineering case. In *Parallel, Distributed and Network-Based Processing (PDP), 2010 18th Euromicro International Conference on*, pages 607–614, 2010.

[7] David Heath, Robert Jarrow, and Andrew Morton. Bond pricing and the term structure of interest rates: A discrete time approximation. *Journal of Financial and Quantitative Analysis*, 25(04):419–440, December 1990.

[8] David Heath, Robert Jarrow, and Andrew Morton. Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica*, 60(1):77–105, January 1992.

[9] Gareth W. Morris and Matthew Aubury. Design space exploration of the european option benchmark using hyperstreams. In *International Conference on Field Programmable Logic and Applications, Amsterdam, The Netherlands*, 2007.

[10] Martingale Musiela. *Methods in Financial Modeling*. Springer-Verlag, 2 Ed., 2004.

[11] Michael J. Quinn. *Parallel Programming in C with MPI and OpenMP*. McGraw-Hill Education Group, 2003.

[12] Xiang Tian and Khaled Benkrid. High-performance quasi-monte carlo financial simulation: Fpga vs. gpp vs. gpu. *ACM Trans. Reconfigurable Technol. Syst.*, 3(4):26:1–26:22, November 2010.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Improved methodology for high-quantiles (VaR) estimator.

J. del Castillo[1], M. Padilla[1] and I. Serra[1]

[1] *Department of Mathematics, Universitat Autònoma de Barcelona*

emails: castillo@mat.uab.cat, mpadilla@mat.uab.cat, iserra@mat.uab.cat

Abstract

The estimation of high quantile is a typical requirement in many areas of application such as insurance and finance. Techniques based on Peaks-over-Threshold (PoT) with parametric and non-parametric novel methods based on residual coefficient of variation. This improvement is applicable at least to market risk, since the existence of three moments can be assumed in market risk.

Key words: Risk analysis, High quantile, Value-at-Risk, Tail index, Heavy tails, Peaks over Thresholds

1 Introduction

In financial mathematics and financial risk management, the value at risk (VaR) is a widely used in Advanced Measurement Approach (AMA) for risk measure. For instance, on a specific portfolio of financial assets, the VaR at level of risk α shows that there is a probability α that the portfolio will fall in value by more than the VaR over a one day period if there is no trading. The extreme value theory (EVT) has two main approaches: Block maxima models and Threshold exceedance models. The first uses as parametric model the generalized extreme value distribution (GEV) and the second the generalized Pareto distribution (GPD). The financial markets provide many data sets where the two approaches may be compared estimating high quantile. Some semi-parametric models based on bias reduction techniques for heavy tails trough the use of an adequate bias-corrected tail index estimator are considered. A new non-parametric tool based on the residual coefficient of variation is also analyzed. This paper focuses on value-at-risk for log-returns arising in modeling extremes of four data sets in the field of finance, widely documented and studied that can be considered with heavy-tail. Applying extreme value statistics in finance requires accurate

 ISBN: 978-84-616-9216-3

estimators on extreme value indices that can be around zero. New parametric models can still be of high interest for the analysis of extreme events, if associated with appropriate statistical inference methodologies, for instance, the full-tails gamma distribution. Remark that, from computational point of view, improved the VaR estimation requires advanced methodologies for high quantile estimation and the current methods consists in to consider statistical models for the extreme values.

2 Techniques for VaR estimation

The main challenge in EVT is to compute the tail index, ν and the optimal threshold, u needed in PoT methodologies. Remark that, some parametric models considered detects the tail index with shape parameter, denoted by $\xi = -1/\nu$. The methodology to model extreme values uses PoT, it is based in the theorem of Pickands-Balkema-DeHaan, see McNeil, *et al.* (2005). From this result, this methodology is used by many authors for modeling exceedances in several fields such as finance and environmental science, for instance Coles (2001). Several techniques have been developed to search for the optimal threshold to link a GPD, such as Hill-plot or ME-plot. This theoretical methodology shows some surprises in practical applications. For instance, Dutta and Perry (2006) observed, in an empirical analysis of operational risk, that even when Pareto distribution fit the data it may result in unrealistic capital estimates (sometimes more than 100% of the asset size). To contribute for a solution to these problems it is necessary to use other alternative models to the GPD, but it requires certain properties that allow them to be treated as queuing models, it is the case of FTG, see Castillo et al. (2012).

The probability density function of the full-tails gamma (FTG) is given by

$$f(x;\nu,\sigma,\theta) = \theta^{\nu}(x+\sigma)^{\nu-1}\exp(-\theta(x+\sigma))/\Gamma(\nu,\sigma\theta) \tag{1}$$

where $\Gamma(\nu,\rho)$ is the upper incomplete gamma function, see Abramowitz y Stegun (1972), the range of x is $(0,\infty)$ and $\nu \in \mathbb{R}, \theta > 0, \sigma > 0$. The tail index is ν, in fact the value of $\xi = -1/\nu$. Remark that for σ fixed, if θ tends to zero, the FTG distribution corresponds to Pareto distribution. The reason that FTG is most appropriate is that the financial data have heavy tails but they have some moments, see Shyriaev (1999). The existence of at least three moments allows us to develop new techniques for extreme values more satisfactory in practice. Furthermore, it should be consider the exponential tails as a first hypothesis, see Castillo et al. (2014).

A new non-parametric tool based on the residual coefficient of variation is described below. This method is applied to the case of generalized Pareto distribution (GPD). Let X be a continuous non-negative random variable (r.v.) with distribution function $F(x)$. For any threshold, $t > 0$, the r.v. of the conditional distribution of threshold exceedances $X - t$ given $X > t$, denoted by $X_t = (X - t \mid X > t)$, is called the *residual distribution* of X over

 ISBN: 978-84-616-9216-3

t. The quantity $M(t) = E(X_t)$ is called the *residual mean* and $V(t) = var\,(X_t)$the *residual variance*. The *residual coefficient of variation* is given by $CV(t) \equiv CV(X_t) = \sqrt{V(t)}/M(t)$, like the usual CV, the function $CV(t)$ is independent of scale. If $CV(t)$ is constant then the distribution of X is a GPD, see Gupta and Kirmani (2000). Remark that, the residual CV for GPD, provided $\xi < 1/2$, is a constant given by $CV^2\,(t) = 1/(1-2\xi)$.

The coefficient of variation can be used also as a measure of non normality. The most popular measure of non normality nowadays is the kurtosis, defined for distributions with four finite moments. The next Proposition shows that the kurtosis can be obtained with the coefficient of variation.

Proposition 1 *Given a symmetric random variate x with respect to zero, the excess kurtosis is*

$$ku\,[x] + 3 = \frac{E\left[x^4\right]}{E\left[x^2\right]^2} = 1 + cv[x^2]^2,$$

therefore the kurtosis is a function of coefficient of variation of x^2.

Finally, the non-parametric CV methodology to compute a tail index estimation corresponds to a computational approach based on to search the value of coefficient of variation that minimizes the distance between its confidence interval under hypothesis of constant tail index and the CV-plot. This non-parametric methodology provides both the tail index and the optimal threshold This methodology combined with Pareto as the model for the tail is denoted by CVm and some examples are showed in Table 3. The last methodology considered is denoted by GPm and it consists in a semi-parametric method for high quantiles estimation based on the parametric model from Pareto and with a non-parametric techniques of bias-corrected Hill estimator, see Gomes and Pestana (2007).

3 Financial data analysis

To compare the different techniques four sets of finance data are considered, collected over the same period: from January 4,1999 through November 17,2005. Those sets of data were the Euro-USA dollar (EUSD) daily exchange rates and the daily closing values of the Dow Jones Industrial Average In (DJI), Microsoft Corp. (MSFT), and International Business Machines Corp. (IBM) stocks. The assumption that financial data have heavy tail can lead to conclusions far removed from reality, in Figure 1 the CV-plot of EUSD shows that the shape parameter can be negative (residual CV less than 1), so a heavy tail it isn't the best option.

The Table 3 shows an brief of the results of the study. The cases GPD and, FTG corresponds to model the whole data as the corresponding parametric model and GEV to model the month maximums. In front of to consider the new methodology CVm and the alternative GPm. DJI and EUSD data analysis shows that the tail of data is not a heavy

 ISBN: 978-84-616-9216-3

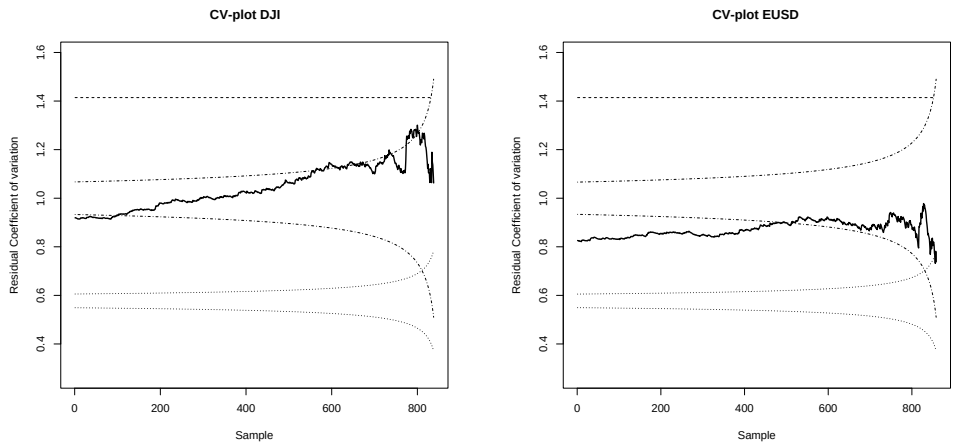


Figure 1: CV-plot of the absolute value of negative tail of log-returns for DJI and EUSD data. Dotdash and dotted line correspond to the 95% confidence interval of a exponential and uniform distribution, respectively.

		99,9%	ξ
DJI	GPD	0,055	0,00
	FTG	0,050	-0,60
	CVm	0,040	0,04
	GPm	1,917	0,30

		99,9%	ξ
EUSD	GPD	0,033	0,00
	FTG	0,026	-0,17
	CVm	0,016	-0,16
	GPm	1,172	0,25

Table 1: A high quantile, shape value ξ for some methodology and data sets: DJI and EUSD.

tail, in fact, the results suggests that the tail distribution has available all the moments. IBM and MSFT data have splits and they have been worked with and without them. Most interesting results, from applied point of view, are obtained using POT with this advanced methodologies to search optimal threshold and improved parametric models for tails as the FTG.

4 Conclusions

After analyzing the data set of this study the following conclusions emerge. Given that EVT is very sensitive to outliers one must be very careful to analyze market data. In practical applications it is recommended to consider the data from different points of view and not be limited to a single technique. The market data, once corrected for splits, is well fitted by models with semi-heavy tails that has few finite moments, as certain authors

 ISBN: 978-84-616-9216-3

claim. When evaluating risks, it is better to study separately the positive and negative tails of the distribution and not doing it together. Thus the coefficient of variation is a more appropriate tool than the kurtosis to assess the weight of the tails.

Acknowledgement

This work has been (partially) supported by Ministerio de Educación y Ciencia (MEC) of Spain under Grant Procesos Estocásticos Aplicados, MTM 2012-31118.

References

[1] Castillo, J. D., Daoudi, J. and Lockhart, R. (2014), Methods to Distinguish Between Polynomial and Exponential Tails. Scandinavian Journal of Statistics, 41: 382–393. doi: 10.1111/sjos.12037.

[2] Castillo, J.D., Daoudi, J. and Serra, I (2012). The full-tails gamma distribution applied to model extreme values. arXiv preprint arXiv:1211.0130.

[3] Coles, S. (2001). An Introduction to statistical of Extremes Values. Springer, London.

[4] Degen, M., Embrechts, P. (2008): EVT-based estimation of risk capital and convergence of high quantiles. Advances in Applied Probability 40(3), 696-715.

[5] Dutta, Kabir, and Jason Perry (2006). A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital.

[6] Embrechts, P., Klüperberg, C. and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer, Berlin.

[7] Finkenstadt, Barbel, and Holger Rootzen (2003). Extreme values in Finance. Telecommunications, and the Environment, Champan & Hall/CRC, Boca Raton, Florida.

[8] Gomes, M. Ivette, and Dinis Pestana (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. Journal of the American Statistical Association 102.477.

[9] Gupta, R. and Kirmani, S. (2000). Residual coefficient of variation and some characterization results. *J. Stat. Plan. Infer.* 91, 23–31.

[10] McNeil, A.J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton Series in Finance.

[11] Shiryaev, Albert N (1999). Essentials of stochastic finance: facts, models, theory. Vol. 23. Singapore: World scientific.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Malcev algebras and combinatorial structures

Manuel Ceballos[1], Juan Núñez[1] and Ángel F. Tenorio[2]

[1] *Departamento 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

In this work, we design an algorithmic method to associate combinatorial structures with finite-dimensional Malcev algebras. In addition to its theoretical study, we have performed the implementation of procedures to construct the digraph associated with a given Malcev algebra (if its associated combinatorial structure is a digraph) and, conversely, a second procedure to test if a given digraph is associated with some Malcev algebra.

Key words: Digraph, Combinatorial structure, Malcev algebra, Combinatorial operations, Algorithm.
MSC 2000: 17D10, 05C25, 05C20, 05C85, 05C90, 68W30, 68R10.

1 Introduction

Research on non-associative algebras is very extensive due to both its own theoretical relevance and its applications to many different fields, like Engineering, Physics or Applied Mathematics. Within these algebras, we will study Malcev algebras. These algebras were introduced by A. I. Malcev [6] as tangent algebras of analytic Moufang loops. They are related to alternative algebras in the same way that Lie algebras are related to associative algebras: if A is an alternative algebra, then the algebra A^- with the operator $[a, b] = ab-ba$ is a Malcev one. However, many general questions about these algebras have not been solved at present by means of traditional techniques, such as obtaining the classification of Malcev algebras.

Currently, Graph Theory has become an essential tool to solve a wide range of problems in different research fields. In this way, we think that graphs and simplicial complexes (its generalization to higher dimensions) may be used to study non-associative algebras and solve open problems like the above-mentioned problem of classifying Malcev algebras.

The main goal of this paper is to start studying the link between combinatorial structures and Malcev algebras. More concretely, we pursue the generalization of the research started in [1] and developed in [2, 3, 4] to the case of Malcev algebras instead of considering Lie algebras.

2 Preliminaries

For a general overview on Malcev algebras and Graph Theory, the reader can consult [7, 5]. We only consider finite-dimensional Malcev algebras over the complex number field $\mathbb{C}$.

Definition 1 *A* Malcev algebra $\mathfrak{g}$ *is a vector space with a second bilinear inner composition law* ($[\cdot,\cdot]$) *called the* bracket product *or* commutator*, which satisfies*

1. $[X,Y] = -[Y,X]$, $\forall X \in \mathfrak{g}$*; and*
2. $[[X,Y],[X,Z]] = [[[X,Y],Z],X] + [[[Y,Z],X],X] + [[[Z,X],X],Y]$, $\forall X,Y,Z \in \mathfrak{g}$.

The second constraint is named the Malcev identity*.*

Given a basis $\{e_i\}_{i=1}^n$ *of* $\mathfrak{g}$*, its* structure (or Maurer-Cartan) constants *are defined by* $[e_i,e_j] = \sum c_{i,j}^h e_h$*, for* $1 \leq i < j \leq n$*.*

Note 1 *Since we are considering a field of characteristic different from* 2*, the first constraint in Definition 1 is equivalent to* $[X,X] = 0, \forall X \in \mathfrak{g}$*.*

Definition 2 *Given a Malcev algebra* $\mathfrak{g}$*, its* center *is* $Z(\mathfrak{g}) = \{X \in \mathfrak{g} \mid [X,Y] = 0,\ \forall Y \in \mathfrak{g}\}$*.*

Definition 3 *A* graph *is a ordered pair* $G = (V,E)$*, where* V *is a non-empty set of vertices and* E *is a set of unordered pairs (edges) of two vertices. If the edges are ordered pairs of vertices, then the graph is named* digraph*.*

3 Associating combinatorial structures with Malcev algebras

Let $\mathfrak{g}$ be an n-dimensional Malcev algebra with basis $\mathcal{B} = \{e_i\}_{i=1}^n$. The structure constants are given by $[e_i,e_j] = \sum_{k=1}^n c_{i,j}^k e_k$. In virtue of the skew-symmetry of the bracket product and Note 1, the pair $(\mathfrak{g},\mathcal{B})$ can be associated with a combinatorial structure built according to the following steps, which are similar to those introduced in [1]

 ISBN: 978-84-616-9216-3

a) Draw vertex i for each $e_i \in \mathcal{B}$.

b) Given three vertices $i < j < k$, draw the full triangle ijk if and only if $(c_{i,j}^k, c_{j,k}^i, c_{i,k}^j) \neq (0,0,0)$. Then, the edges ij, jk and ik have weights $c_{i,j}^k$, $c_{j,k}^i$ and $c_{i,k}^j$, respectively.

   b1) Use a discontinuous line (named *ghost edge*) for edges with weight zero.

   b2) If two triangles ijk and ijl with $1 \leq i < j < k < l \leq n$ satisfy $c_{i,j}^k = c_{i,j}^l$, draw only one edge between vertices i and j shared by both triangles; see Figure 1.

c) Given two vertices i and j with $1 \leq i < j \leq n$ and such that $c_{i,j}^i \neq 0$ (resp. $c_{i,j}^j \neq 0$), draw a directed edge from j to i (resp. from i to j), as can be seen in Figure 2.

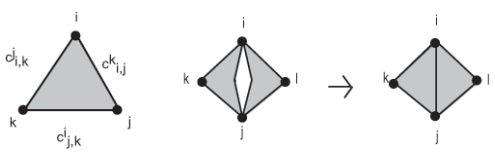


Figure 1: Full triangle and two triangles sharing an edge.

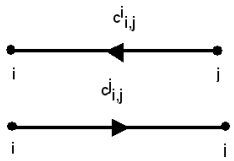


Figure 2: Directed edges.

4 Theoretical results

This section is devoted to state some general results on the association between Malcev algebras and combinatorial structures. We start considering some general properties arising from this association and corresponding to topological properties of the combinatorial structure.

Proposition 1 *Let G be the combinatorial structure associated with a Malcev algebra $\mathfrak{g}$. If v is an isolated vertex of G, then the basis vector $e_v \in \mathfrak{g}$ associated with v belongs to the center $Z(\mathfrak{g})$.*

Proposition 2 *Let G be the combinatorial structure associated with a Malcev algebra $\mathfrak{g}$. Each connected component of G is associated with a Malcev subalgebra of $\mathfrak{g}$. Moreover, if G is non-connected, then $\mathfrak{g}$ is the direct sum of the Malcev subalgebras associated with the connected components of G.*

 ISBN: 978-84-616-9216-3

Next, we have studied the particular case in which there are no full triangles in the combinatorial structure (i.e. a weighted digraph). Let us note that this assertion is equivalent to consider a Malcev algebra $\mathfrak{g}$ with basis $\mathcal{B} = \{e_i\}_{i=1}^n$ and law

$$[e_i, e_j] = c_{i,j}^i e_i + c_{i,j}^j e_j,\ 1 \leq i < j \leq n. \tag{1}$$

Proposition 3 *If G is a connected digraph with 3 vertices associated with a Malcev algebra, then G must be isomorphic to some of the configurations shown in Figure 3.*

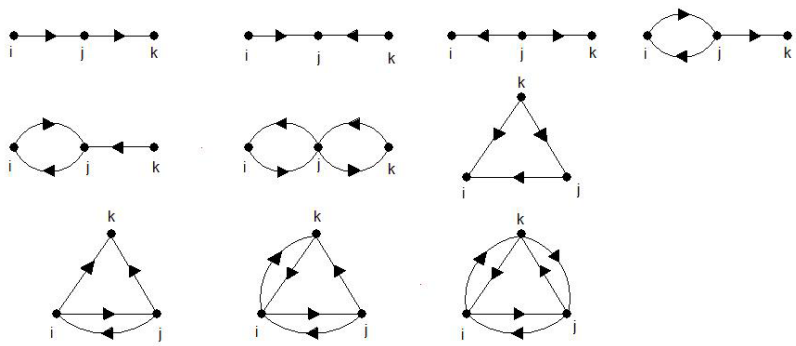


Figure 3: Connected digraphs with 3 vertices associated with Malcev algebras.

Note 2 *Some of the configurations shown in Figure 3 require particular restrictions on the structure coefficients in order to assure its association with Malcev algebras.*

Corollary 1 *The connected digraphs with 3 vertices shown in Figure 4 cannot be contained in a digraph associated with a Malcev algebra of any given dimension (i.e. they are forbidden configurations).*

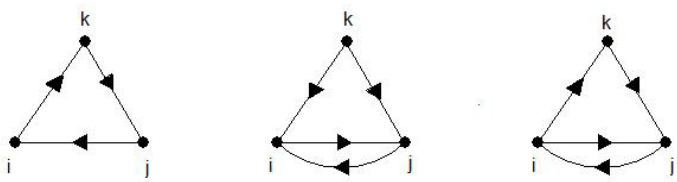


Figure 4: Forbidden configurations in digraphs associated with Malcev algebras.

 ISBN: 978-84-616-9216-3

5 Algorithmic procedures

In this section we present two algorithms dealing with converse questions: the first is devoted to obtain the digraph associated with a given Malcev algebra starting from its law; whereas the second provides a test to determine if a weighted digraph is associated with a Malcev algebra or not.

5.1 Algorithm to obtain the digraph associated with a Lie algebra

Under the same notation as in Section 4, we consider an n-dimensional Malcev algebra $\mathfrak{g}$ with basis $\mathcal{B}_n$. Hence, we are considering a law consisting only of brackets $[e_i, e_j] = c_{i,j}^i e_i + c_{i,j}^j e_j$ to avoid full triangles and deal only with digraphs.

We have designed the following algorithm to obtain the digraph associated with $\mathfrak{g}$, structured in four steps

1. Computing the bracket product between two arbitrary basis vectors in $\mathcal{B}$.
2. Evaluating the bracket between two vectors expressed as a linear combination of vectors from basis $\mathcal{B}$.
3. Imposing the Malcev identity and solving the corresponding system of equations.
4. Drawing the digraph associated with the Malcev algebra $\mathfrak{g}$.

To implement the algorithm, we use 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 algebra $\mathfrak{g}$.

5.2 Algorithm to decide if a digraph is associated with a Malcev algebra

We show an algorithmic procedure to determine if a given digraph is associated or not with a Malcev algebra. The algorithm consists of the following two steps

a) Generating the law candidate to be a Malcev algebra using in reverse the construction in Section 3.

b) Checking if the Malcev identities are satisfied for this law.

To implement the algorithm, we need load the libraries `GraphTheory` and `DifferentialGeometry`. The first library activates commands related to Graph Theory; whereas the second provides

 ISBN: 978-84-616-9216-3

some simplifications to translate the digraph in a vector space with a bilinear bracket product.

More concretely, we start defining the vector space associated with the digraph by using a routine which receives as inputs the list `V` with the vertices of the digraph and the set `E` with its directed, weighted edges. As outputs, 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 non-zero brackets coming from the weighted edges in the set `E`.

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 Appl. 389 (2004), 43–61.

[2] J. Cáceres, M. Ceballos, J. Núñez, M.L. Puertas and A.F. Tenorio, *Combinatorial structures of three vertices and Lie algebras*, Int. J. Comput. Math. 89 (2012), 1879–1900.

[3] M. Ceballos, J. Núñez and A.F. Tenorio, *Complete triangular structures and Lie algebras*, Int. J. Comput. Math. 88 (2011), 1839–1851.

[4] M. Ceballos, J. Núñez and A.F. Tenorio, *Study of Lie algebras by using combinatorial structures*, Linear Algebra Appl. 436:2 (2011), 349–363.

[5] F. Harary, *Graph Theory*, Addison-Wesley, Reading, 1969.

[6] A.I. Malcev, *Analytic loops*, Mat. Sb. 78 (1955), 569–578.

[7] A.A. Sagle, *Malcev Algebras*, Trans. Amer. Math. Soc. 101 (1961), 426–458.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7July, 2014.

Accelerating HEVC using GPU-based heterogeneous platforms

Gabriel Cebrián-Márquez[1], José Luis Martínez[1], Pedro Cuenca[1], Minhao Tang[2] and Jiangtao Wen[2]

[1] *Albacete Research Institute of Informatics (I3A), University of Castilla-La Mancha (Albacete, SPAIN)*

[2] *Department of E.E, Tsinghua University (Beijing, CHINA)*

emails: Gabriel.Cebrian@uclm.es, JoseLuis.Martinez@uclm.es, Pedro.Cuenca@uclm.es, tmh920811@163.com, jtwen@tsinghua.edu.cn

Abstract

HEVC standard achieves double compression efficiency compared with H.264/AVC at the cost of a huge computational complexity. Parallelizing HEVC encoding is an efficient approach to fulfil this high computational requirement. The parallelization approaches considered in HEVC (such as Slices, Tiles, WPP) rely on creating picture partitions that can be processed concurrently. This paper focuses on the realization of a parallel architecture design of heterogeneous platforms composed by a GPU plus a multi-core CPU to take advantages of these techniques. Experimental results outperform WPP in terms of speed-up and coding efficiency. Moreover, the proposed parallel method obtains an overall speed-up of more than 4x in an Intel quad-core CPU and a NVIDIA GPU with negligible quality loss compared to the non-parallel version.

Key words: HEVC, Parallelization, GPU, Multicore, heterogeneous computing

1 Introduction

Recently, the new *High Efficiency Video Coding* (HEVC) standard [1] has been established by the *Joint Collaborative Team on Video Coding* (JCT-VC), an expert group proposed by the ISO/IEC Moving Expert Group (MPEG) and ITU-T Video Coding Expert Group (VCEG). HEVC was initially conceived with the purpose of achieving a highly efficient performance for delivering high quality multimedia services over bandwidth-constrained

networks, but also to give support to formats beyond HD resolution, such as the new 4K and 8K formats. This standard is based on a well-known block-based hybrid video coding architecture as well as its predecessor H.264/MPEG4 part 10 - Advanced Video Coding (AVC) [2], which it outperforms in terms of bitrate reduction at the same quality [3]. Among others, HEVC includes multiple new coding tools, namely highly flexible quad-tree coding block partitioning which includes new concepts as *Coding Unit* (CU), *Prediction Unit* (PU) and *Transform Unit* (TU) [3, 4].

All these improvements imply a considerable increase of the encoding time. Fortunately, this computational cost can be efficiently reduced by adapting the sequential encoding algorithm to parallel architectures. Over the last few years the computation industry has tended towards including several processing units in a single shared chip. Furthermore, in terms of massive data computations, there are also devices called *Graphic Processing Units* (GPUs). These devices, also referred as many-core, are highly parallel and they are normally used as co-processors to assist the *Central Processing Unit* (CPU). CPUs and GPUs have different instruction set architectures, forming what it is known as a heterogeneous computing platform [5].

As a support to this parallelism, HEVC addresses a special emphasis on a hardware friendly design and parallel-processing architectures. These parallelization approaches are *Tiles* [6] and *Wavefront Parallel Processing* (WPP) [7] that will be depicted in Section 2. Basically, these parallelization approaches rely on creating picture partitions that break some dependencies for prediction, CABAC context modelling, and/or slice header overhead. As a result, coding losses may appear.

At this point, this paper proposes a GPU-based algorithm that makes use of this device in order to efficiently parallelize the motion estimation carried out in the HEVC inter-prediction algorithm. Furthermore, this algorithm can be combined, in turn, with multiple coarse-grained parallelization techniques such as the aforementioned ones in a heterogeneous architecture. In fact, this paper shows the results provided by a combination of the WPP algorithm and this GPU-based proposal.

These two algorithms are tested comparing their results with the ones provided by the HEVC Test Model (HM) [8], outperforming them in terms of speed-up and coding efficiency; moreover, compared with the sequential version of HEVC, speed-up is increased up to 4.53x in a quad-core CPU (4 threads plus SMT) with negligible *Rate Distortion* (RD) penalty.

The remainder of this paper is organized as follows: Section 2 includes a technical background of the new HEVC standard while Section 3 identifies the related work which is being developed about the topic. Section 4 introduces our proposed architecture. Experimental results are shown in Section 5. Section 6 concludes the paper and includes some lines of action as future work.

 ISBN: 978-84-616-9216-3

2 Technical background

As mentioned in the previous section, the main target of HEVC is to achieve lower bitrates for video streams while maintaining the same quality. In order to make this possible, HEVC introduces new coding tools with respect to its predecessor, H.264/AVC; all of them make it possible to notably increase coding efficiency. One of the most important changes affects the picture partitioning. HEVC dispenses with the terms *Macro-Block* (MB) and *Block* for the ME and the transform, respectively, and introduces three new concepts: CU, PU and TU. This structure leads to a flexible coding to suit the particularities of the frame. Each picture is partitioned into square regions of variable size called CUs, which replace the MB structure of previous standards. Each CU, whose size is limited from 8x8 to 64x64 pixels, may contain one or several PUs and TUs. To fix the size of each CU, first of all a picture is divided into 64x64 pixels areas, which are called *Coding Tree Units* (CTU), and then, each CTU can be partitioned into 4 smaller sub-areas of a quarter of the original area. This partitioning can be performed with each sub-area recursively until it has a size of 8x8 pixels, as shown in Figure 1.

For intra-picture prediction, a PU uses the same 2Nx2N size as of the CU to which it belongs, allowing it to be split into quad NxN PUs only for CUs at the minimum depth level. Therefore the PU size ranges from 64x64 to 4x4 pixels. For inter-picture prediction, several non-square rectangular block shapes are available in addition to square ones, allowing eight different PU sizes (2Nx2N, 2NxN, nx2N, NxN, 2NxU, 2NxD, nLx2N, nRx2N). The prediction residual obtained in each of the PUs is transformed using the *Residual Quad Tree* (RQT) structure, which supports various TU sizes from 32x32 to 4x4. For the transform coding of intra prediction 4x4 PU residuals, an integer approximation of the *Discrete Sine Transform* (DST) is used instead.

HEVC checks most of the PUs (Inter and Intra modes) to decide whether it should split a CU or not by choosing the best RD case. Furthermore, in the case of Inter Prediction, for each of these PU partitions an ME algorithm is called. This wide range of possibilities makes HEVC much more computationally expensive than its predecessor, H.264/AVC. HEVC introduces changes in other modules too, such as Intra Prediction (where a total of 35 different coding modes can be selected), the PU modes (it introduces asymmetric modes), new image filters or new transform sizes, among others. As expected, the selection of the optimal partitioning for each CU/PU/TU is an intensive time-consuming process due to the huge number of combinations that have to be evaluated in order to achieve the best performance.

With the aim of reducing this huge complexity, the new HEVC codec also includes new parallelization techniques such as tiles [6] and WPP [7] among slices. On the one hand, tiles are square or rectangular shape partitions where dependencies are broken across tile boundaries [6], making it possible to process them independently, taking into account that coding losses may appear. The in-loop filters (deblocking and SAO), however, can still

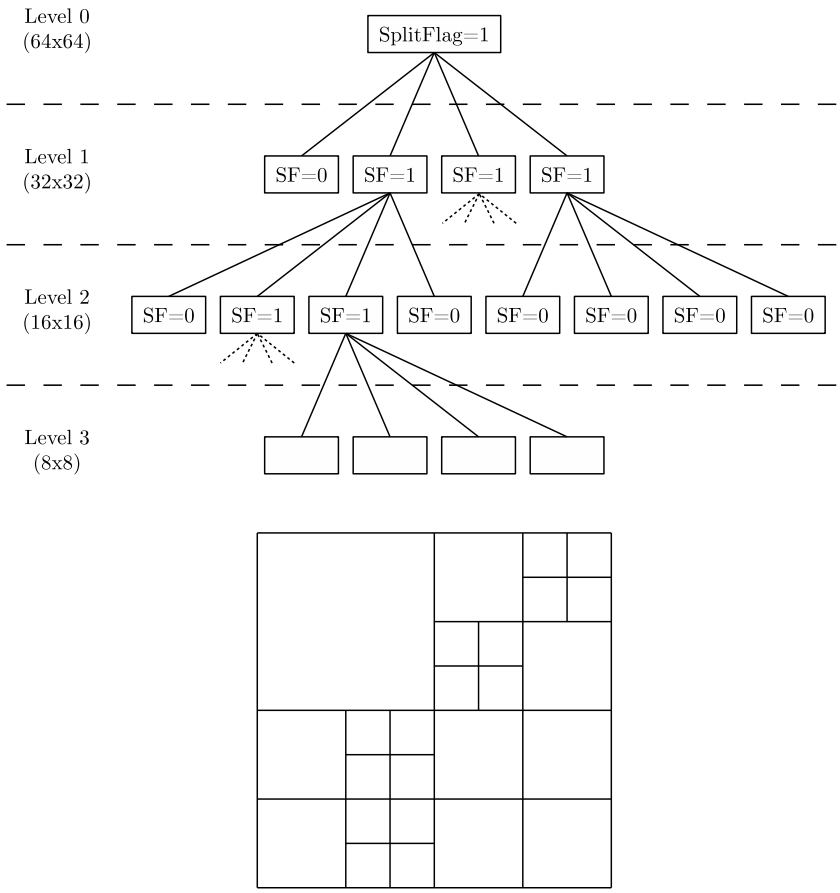
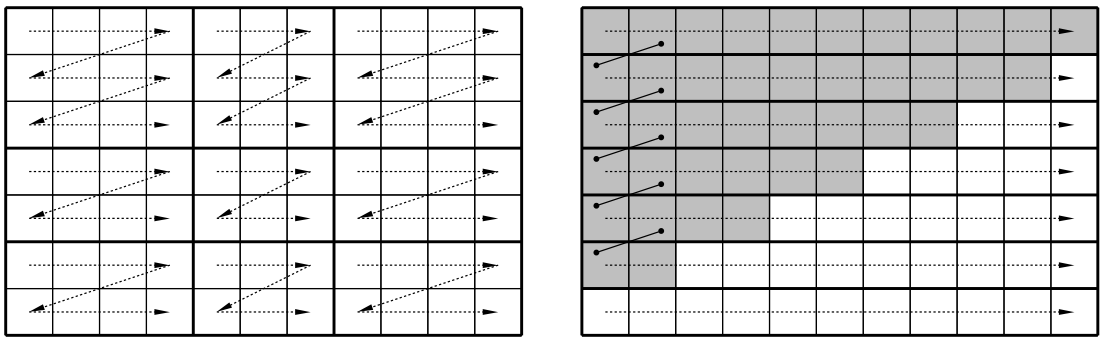


Figure 1: CTU quadtree structure partitioning.

cross these boundaries. The number of tiles and their location can be defined for the entire sequence or changed from picture to picture. On the other hand, WPP allows the creation of picture partitions (normally rows) that can be processed in parallel, whereas entropy encoding and prediction are allowed to cross partitions in order to minimize coding losses. Nevertheless, coding dependencies make it necessary to have a delay of at least two CUs between consecutive rows in a similar way as segmentation does in a computer architecture [7, 9]. For this reason, not all the processes can start encoding these rows at the same time, which involves a low CPU utilization at the beginning and at the end of a frame, incurring in the so-called "ramping inefficiencies". Both techniques are depicted in Figure 2.

Tiles and WPP have different merits and disadvantages. While WPP is generally well suited for the parallelization of the encoder and the decoder due to its high number of picture partitions with low compression losses, the amount of parallelism with tiles is not

 ISBN: 978-84-616-9216-3



(a) Tiles partitioning (b) WPP partitioning

Figure 2: Partitioning and processing order of tiles (a) and WPP (b)

fixed, as the number of regions in which a frame is divided may vary. Additionally, WPP does not introduce artifacts at partition boundaries as is the case for Tiles. In order to simplify the implementation, it is not possible to use Tiles and WPP simultaneously in the same compressed video sequence.

In either case, these approaches need parallel architectures to exploit their potential and, hence, reduce the computational complexity of HEVC. In this respect, new architectures are being introduced in high-performance computing composed of multi-core CPUs and GPUs. A multi-core processor is composed of several processors sharing the same chip, while GPUs are composed of hundreds of similar simple processing cores which are designed and organized with the goal of achieving high performance. These cores are grouped in stream processors that perform *Single Instruction Multiple Data* (SIMD) operations which are suitable for arithmetic intensive applications. In the particular case of NVIDIA, a powerful GPU architecture called the *Compute Unified Device Architecture* (CUDA) [10] has been developed. 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.

3 Related work

As far as related work in the literature is concerned, in the past, there have been many approaches focusing on accelerating different modules of the H.264/AVC encoding algorithm by means of parallel computing [11, 12, 13]. On the contrary, in the framework of HEVC, the first parallel approaches were focused on reducing the complexity of the decoding algorithm; in [9], the authors improve the WPP approach included in the HM reference software [8]. The idea consists of once there are no available rows in the current picture, the next one starts being processed. In this way, the ramping inefficiencies of WPP can be mitigated

by overlapping the execution of consecutive pictures. This proposal was called *Overlapped Wavefront* (OWP). As a limitation, search areas need to be constrained to the region of the reference frame that has been already reconstructed.

In the context of this paper, which is more focused on the encoder side, there are not many approaches. OWF might work for the encoder, but no results were given in [9]. *Yu et al.* proposed in [14] a parallel candidate list in order to parallelize the motion vector prediction, but the proposal is not standard compliant. Later, in [15], the authors reduced the encoding time up to 13 times by using a 64-core architecture, which is far more expensive than the one used in this paper. Finally, *Wang et al.* proposed in [16] a scheme similar to the one proposed in this paper based on a GPU plus multi-core CPU, but the major lack of this paper lies in the fact that they did not use the reference software HM [8] and, thus, the RD results are worse due to the fact that not all coding tools were implemented [16].

4 Proposed algorithms

As seen before, parallelization is possible in both the encoder and the decoder by using the algorithms defined in the standard. Nonetheless, these are designed to be executed in a multi-core CPU, taking advantage of the capabilities that multiple threads may offer, but not taking into account other devices. Heterogeneous architectures such as the ones formed by the association of a multi-core CPU and a GPU are utilized in this paper, making use of the immeasurable power they can provide. The joint of a GPU-based motion estimation algorithm and the standard WPP is proposed in the following subsections.

4.1 GPU-based inter prediction algorithm

As motion estimation is the most resource intensive operation on the encoder side [3], this algorithm aims to reduce the time spent on the CPU by performing these searches on a GPU device. Nevertheless, taking into account that data transfers between host and GPU are highly time-consuming, these operations are performed asynchronously. In this way, time spent on *Integer Motion Estimation* (IME) is negligible compared with the default search algorithm.

As soon as a *Group of Pictures* (GOP) starts being processed, it is possible to transfer the original frames that will be encoded to the device, making them available for subsequent uses. Later on, these frames are updated with their reconstructed version when they are encoded (and decoded in-loop) in order to correctly carry out motion estimation on the device.

When the encoder starts processing a slice, the host queues the execution of two consecutive kernels that perform the integer motion estimation of every *Prediction Unit* (PU) partition in the first *Coding Tree Unit* (CTU). The first kernel executes the required operations to calculate the *Sum of Absolute Differences* (SAD) residuals across a search area in

 ISBN: 978-84-616-9216-3

the reference frame, while the second one determines which one of them may offer the best possible result.

This algorithm relies on the fact that every PU size established by the standard is divisible by four, and taking into account the nature of the SAD operation, it is possible to calculate the residual information of a PU partition from the composition of its 4x4 SAD partitions.

Following this approach, the previously mentioned kernel distributes a device thread per sample in the reference search area. Every thread is responsible for calculating all the 4x4 SAD blocks in a CTU, taking as motion vector its position in the search area. Once these blocks are calculated, all the running threads put them together to obtain the PU partitions in which a CTU might be divided. From another point of view, the results of this step would be equivalent to a full-search algorithm performed for every PU partition.

At this point, the second kernel performs a reduction algorithm over the residual data obtained from the first one, so that the result of the GPU algorithm is an only table containing the best *Motion Vector* (MV) for every PU partition, which is copied asynchronously to the host. After the transfer is finished, motion search operations related to the next CTU are then issued to the device.

By the time the host needs to perform the motion estimation of the CTU, integer MVs should be ready to be queried, only being necessary to perform fractional motion estimation of the PU partitions which have not been skipped by the encoder.

4.2 Joint algorithm: WPP + GPU-based inter prediction

As a consequence of the computational limit of a single processor, the idea of having multiple cores in the same chip was successfully introduced some time ago. One of its most relevant benefits is that a parallel application can achieve speed-up values in direct proportion to the number of cores. This, along with the wide existence of this kind of devices, motivated the JCT-VC to include parallelism in HEVC, which was carried out by breaking some dependencies while trying to provide as much coding efficiency as possible.

In our heterogeneous architecture, both the multi-core CPU and the GPU algorithms are independent. While tiles or WPP perform a coarse-grained parallelization of the whole encoding process, our GPU-based algorithm carries out the IME operation. This independence makes it possible to combine both algorithms in a single proposal, obtaining, hence, higher speed-up values at the expense of a negligible increment in coding efficiency losses.

As WPP can achieve similar speed-ups with regard to tiles [9] without breaking as many dependencies (and hence, obtaining better coding efficiency results), this is the algorithm taken as the basis of our joint algorithm.

As depicted in the Figure 3, a single GPU device can carry out the integer motion estimation of multiple threads and, hence, multiple CTUs. Therefore, it is necessary to queue several kernels into the GPU. In this way, the device is fully utilized, lowering idle

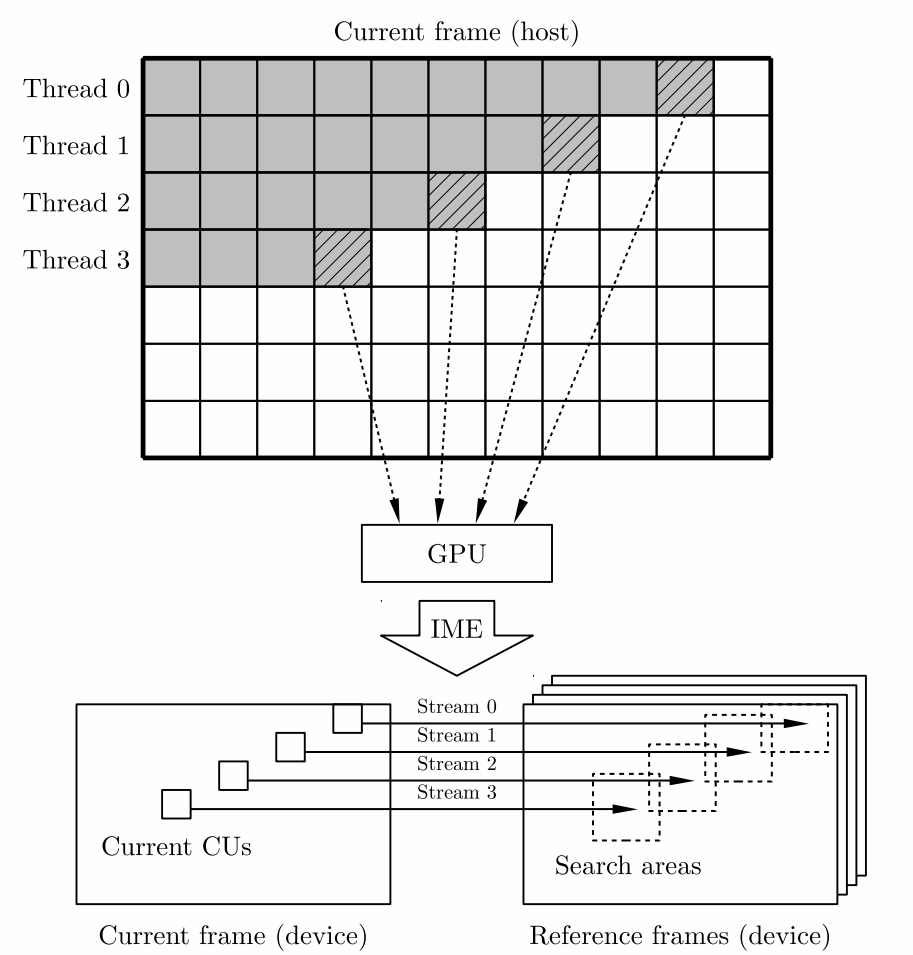


Figure 3: Joint algorithm applied to WPP with 4 threads.

times. In addition, the GPU can process different kernels independent of the CPU, so this one can continue processing other modules concurrently.

5 Performance evaluation

In order to ensure a common framework, JCT-VC defined a document in [17] where test conditions are set out to homogenize comparisons between experiments. Therefore, this performance evaluation has been carried out in accordance with these guidelines.

 ISBN: 978-84-616-9216-3

Table 1: Speed-up and coding efficiency results of the GPU-based proposal.

	Speed-up	BD-rate (%)
Class A	1.12	-0.2
Class B	1.12	0.7
Class C	1.14	0.0
Class D	1.11	-0.2
Class F	1.10	0.9

The sequential algorithm of HM version 10.0 [8] has been used as the reference algorithm to calculate the corresponding speed-up and coding efficiency values. The proposed GPU-based algorithm has been isolatedly tested in order to calculate its influence in the overall processing time. Later, the results of WPP and the joint algorithm are presented.

Random Access has been the chosen configuration to carry out the evaluation, as it is the most widely used configuration in real scenarios, but any other configuration might work with our proposed algorithm as well. No other changes were made to the default parameters provided by the reference software.

All measurements have been performed on a quad-core Intel Core i7-2600 CPU running at 3.40GHz and a NVIDIA GTX 560 Ti GPU running 384 CUDA cores at the frequency of 1.6 GHz. Consequently, tests have been carried out with 2 and 4 threads, as well as 4 plus *Simultaneous MultiThreading* (SMT), enabling the processor to execute 8 threads.

To start with, Table 1 shows the results of the proposed GPU-based algorithm. As can be seen, performing the IME operation on the GPU involves accelerating the encoding process by 1.12x while incurring in very low coding efficiency losses (due to MVs prediction), or even improving it in some cases. This is because the proposed algorithm performs a more exhaustive search. We would like to emphasize that these results are the theoretical limit of the integer ME, as the GPU has already calculated every MV when the host needs to perform ME. In other words, the IME is performed in virtually perfect time.

On the other hand, Table 2 depicts a comparison between the results provided by the joint proposal (WPP + GPU) and the ones provided by WPP itself. Both algorithms have been executed with 2, 4 and 4 plus SMT threads, showing that the proposal can reach speed-up values close to the ones from a parallel efficient algorithm, (i.e. threads are almost fully utilized), providing that the frame size is large enough to exploit the available parallelism.

These results also show that combining both WPP and the GPU-based algorithm surpass the results of WPP in terms of speed-up, reaching values up to 4.33x average (for class A) compared with 3.92x, respectively. Moreover, this increase has a negligible impact of 1.3% BD-rate in terms of coding efficiency. As can be seen, these results can be connected with the ones from Table 1, as the difference in speed-up and BD-rate compared with the GPU-based algorithm itself stands at around 1.10x and 0.3%, respectively (compared with

Table 2: Speed-up and BD-rate results comparison between WPP and our joint proposal.

		Speed-up						BD-rate (%)	
		2 threads		4 threads		4 th. + SMT			
		WPP	Joint	WPP	Joint	WPP	Joint	WPP	Joint
Class A	Traffic	1.88	1.99	3.37	3.56	3.90	4.12	0.7	0.9
	PeopleOnStreet	1.89	2.17	3.35	3.85	3.95	4.53	0.7	0.1
Class B	Kimono	1.89	2.13	3.36	3.80	3.80	4.31	1.2	1.7
	ParkScene	1.88	2.01	3.33	3.56	3.70	3.97	0.7	0.7
	Cactus	1.89	2.07	3.31	3.65	3.72	4.13	1.1	1.4
	BasketballDrive	1.90	2.22	3.40	3.98	3.76	4.43	1.5	5.0
	BQTerrace	1.88	2.00	3.31	3.53	3.79	4.06	1.2	0.1
Class C	BasketballDrill	1.80	2.00	2.73	3.09	2.72	3.09	1.4	1.1
	BQMall	1.81	1.96	2.83	3.11	2.83	3.09	1.5	2.3
	PartyScene	1.78	1.91	2.70	2.94	2.71	2.96	0.6	0.2
	RaceHorses	1.78	2.09	2.79	3.30	2.84	3.35	0.8	0.8
Class D	BasketballPass	1.67	1.88	1.75	2.01	1.75	2.01	0.9	0.9
	BQSquare	1.60	1.66	1.84	1.92	1.84	1.92	1.3	1.3
	BlowingBubbles	1.59	1.66	1.80	1.91	1.80	1.90	0.9	0.6
	RaceHorses	1.63	1.85	1.81	2.10	1.81	2.10	0.9	0.4
Class F	BasketballDrillText	1.79	1.97	2.71	3.05	2.70	3.05	1.4	0.7
	ChinaSpeed	1.86	2.12	3.15	3.54	3.28	3.74	0.8	-2.3
	SlideEditing	1.80	1.86	3.13	3.24	3.33	3.46	1.0	3.9
	SlideShow	1.80	1.92	3.00	3.21	3.24	3.45	2.2	6.7
	Class A	1.88	2.08	3.36	3.71	3.92	4.33	0.7	0.5
	Class B	1.89	2.08	3.34	3.70	3.75	4.18	1.1	1.8
	Class C	1.79	1.99	2.77	3.11	2.78	3.13	1.1	1.1
	Class D	1.62	1.76	1.80	1.99	1.80	1.98	1.0	0.8
	Class F	1.81	1.97	3.00	3.26	3.14	3.43	1.3	2.3
	Average	1.80	1.98	2.85	3.15	3.08	3.41	1.0	1.3

1.12x and 0.2%). This means that the device is almost fully utilized, taking advantage of its potential.

6 Conclusion and future work

In this paper, we have designed an efficient parallel framework of the HEVC encoder on a multi-core CPU plus GPU platform. A coarse-grained parallelization of the whole encoding process is made on a multi-core CPU, while the GPU carries out the ME operation. Comparing our approach to WPP, our experiments show that the proposed joint algorithm achieves better performance in terms of speed-up with negligible coding efficiency losses.

 ISBN: 978-84-616-9216-3

Ongoing work will focus on using multiple GPUs and parallelizing other modules, as well as considering other architectures such as Intel Xeon Phi [18].

Acknowledgements

This work has been jointly supported by the MINECO and European Commission (FEDER funds) under the project TIN2012-38341-C04-04.

References

[1] B. Bross, W. Han, J. Ohm, G. Sullivan, Y.-K. Wang, and T. Wiegand, *High efficiency video coding (HEVC) text specification draft 10*, Doc. JCTVC-L1003, January 2013.

[2] ITU-T and ISO/IEC JTC, *Advanced video coding for generic audiovisual services*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC) version 16, 2012.

[3] J. Ohm, G.J. Sullivan, H. Schwarz, Thiow Keng Tan, and T. Wiegand, *Comparison of the coding efficiency of video coding standards - Including High Efficiency Video Coding (HEVC)*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1669–1684, Dec 2012.

[4] F. Bossen, B. Bross, K. Suhring, and D. Flynn, *HEVC complexity and implementation analysis*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1685–1696, Dec 2012.

[5] Wu chun Feng and Dinesh Manocha, *High-performance computing using accelerators*, Parallel Computing, vol. 33, no. 10-11, pp. 645–647, 2007.

[6] K. Misra, A. Segall, M. Horowitz, Shilin Xu, A. Fuldseth, and Minhua Zhou, *An overview of tiles in HEVC*, IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 6, pp. 969–977, Dec 2013.

[7] F. Henry and S. Pateux, *Wavefront Parallel Processing*, Tech. Rep. JCTVC-E196, March 2011.

[8] *HM reference Software*, https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/.

[9] Chi Ching Chi, M. Alvarez-Mesa, B. Juurlink, G. Clare, F. Henry, S. Pateux, and T. Schierl, *Parallel scalability and efficiency of HEVC parallelization approaches*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1827–1838, Dec 2012.

[10] NVIDIA, *NVIDIA CUDA Compute Unified Device Architecture programming guide, version 3.2*, August 2010.

[11] Ngai-Man Cheung, Xiaopeng Fan, O.C. Au, and Man-Cheung Kung, *Video coding on multicore graphics processors*, Signal Processing Magazine, IEEE, vol. 27, no. 2, pp. 79–89, March 2010.

[12] Chenggang Yan, Feng Dai, and Yongdong Zhang, *Parallel deblocking filter for H.264/AVC on the TILERA many-core systems*, in Advances in Multimedia Modeling, vol. 6523 of Lecture Notes in Computer Science, pp. 51–61. Springer Berlin Heidelberg, 2011.

[13] Huayou Su, Nan Wu, Chunyuan Zhang, Mei Wen, and Ju Ren, *A multilevel parallel intra coding for H.264/AVC based on CUDA*, in Image and Graphics (ICIG), 2011 Sixth International Conference on, Aug 2011, pp. 76–81.

[14] Qin Yu, Liang Zhao, and Siwei Ma, *Parallel AMVP candidate list construction for HEVC*, in Visual Communications and Image Processing (VCIP), 2012 IEEE, Nov 2012, pp. 1–6.

[15] Chenggang Yan, Yongdong Zhang, Feng Dai, and Liang Li, *Highly Parallel Framework for HEVC motion estimation on many-core platform*, in Data Compression Conference (DCC), 2013, March 2013, pp. 63–72.

[16] Xiangwen Wang, Li Song, Min Chen, and Junjie Yang, *Paralleling variable block size motion estimation of HEVC on CPU plus GPU platform*, in IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2013, July 2013, pp. 1–5.

[17] JCT-VC, *Common test conditions and software reference configurations*, Join Collaborative Team on Video Coding 12th meeting, Doc. JCTVC-L1100, January 2013.

[18] Jianbin Fang, Ana Lucia Varbanescu, and Henk Sips, *Identifying the key features of Intel Xeon Phi: a comparative approach*, Parallel and Distributed Systems Report Series, Report number PDS-2013-006, May 2013.

 ISBN: 978-84-616-9216-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7 July, 2014.

Convergence Analysis and Applications of Operator Splitting Methods for Burgers-Huxley Equation

Yesim Çiçek[1] **and Gamze Tanoğlu**[1]

[1] *Department of Mathematics, Izmir Institute of Technology*

emails: yesimyazici@iyte.edu.tr, gamzetanoglu@iyte.edu.tr

Abstract

We provide an error analysis of the operator splitting method of the operator splitting of the Godunov and Strang type applied to the Burgers-Huxley equation, $u_t + \alpha u u_x - \epsilon u_{xx} = \beta(1-u)(u-\gamma)u$. The major task is to prove the convergence rates for the two splitting methods in Sobolev spaces.

We split the equations into linear and nonlinear parts and show that the operator splitting methods have the correct convergence rates in $H^s(\mathbb{R})$, where $H^s(\mathbb{R})$ is the Sobolev space and s is an arbitrary nonnegative integer.

We numerically apply the operator splitting methods to the Burgers-Huxley equation for the split step size Δt.

Key words: Operator splitting, Burgers-Huxley equation, nonlinear.

1 Introduction

In this paper, we study the Burgers-Huxley equation which is given in the following form,

$$u_t + \alpha u u_x - \epsilon u_{xx} = \beta(1-u)(u-\gamma)u, \tag{1}$$

where $x \in \mathbb{R}$, $t > 0$, $\alpha, \beta \geq 0$, $0 < \epsilon \leq 1$ and $0 < \gamma < 1$. There are many numerical methods which have been studied to compute approximate solutions to Burgers-Huxley equation. The idea of operator splitting, (see [2], [3], [6], [7], [8], [9], [10], [11], [13], [14] and [15]), is widely used for the approximation of partial differential equations. The basic idea is based on splitting a complex problem into simpler sub-problems, each of which is solved by an efficient method. One of the reasons for the popularity of operator splitting is the use of dedicated special numerical techniques for each of the equations.

 ISBN: 978-84-616-2723-3

Assume the time $T > 0$ is fixed and consider a general partial differential equation

$$u_t = C(u), \quad t \in [0, T], \quad u(0) = u_0, \tag{2}$$

where $C(u)$ is a differential operator between some normed spaces, say X, and assume u_0 and solution $u(t)$ are in X. We assume that the Taylor series expansion is valid for $u(t)$, which results in

$$u(t) = u(0) + tu_t(0) + \mathcal{O}(t^2). \tag{3}$$

If we replace the second term in the above series with (2) we get

$$u(t) = u(0) + tC(u_0) + \mathcal{O}(t^2). \tag{4}$$

Furthermore, assume $C(u)$ can be written as a sum of more elementary operators, say

$$C(u) = A(u) + B(u), \tag{5}$$

which yield

$$u(t) = u_0 + t(A(u_0) + B(u_0)) + \mathcal{O}(t^2). \tag{6}$$

The operator splitting method is built up as follows: Fix a positive and small time step Δt, and discretize the time with n steps such that $t_n \leq n\Delta t$. Instead of solving equation (2) directly, we solve the two subequations

$$\begin{aligned} v_t &= A(v) \\ w_t &= B(t), \end{aligned} \tag{7}$$

for each time step, and concantenate the solutions. The simplest form for an operator splitting solution of (2) is formed solving the first subequation using the solution from the second subequation as initial condition when solving at each time step. Writing out this procedure gives,

$$u_{n+1} = e^{A\Delta t}(e^{B\Delta t}(u_n)) = e^{A\Delta t} \circ e^{B\Delta t}(u_n) = [e^{A\Delta t} \circ e^{B\Delta t}]^n(u_0), \tag{8}$$

where u_n is the operator splitting solution at time t_n, and $e^{At}(v_0)$ and $e^{Bt}(w_0)$ are the exact solution operators of the above subequations at time t with initial data v_0 and w_0. This is the well-known *Lie-Trotter splitting method*.

Other and more sophisticated methods for an operator splitting solution of (2) are created by solving two subequations for different split step sizes, and compose the solution operators in a more complicated way. By solving one of subequations for half the step size composed with the solution of the other subequation for a full time step, we obtain the famous *Strang splitting method*, which is given as

$$\begin{aligned} u_{n+1} &= e^{A\Delta t/2}(e^{B\Delta t}(e^{A\Delta t/2}(u_n))) \\ &= e^{A\Delta t/2} \circ e^{B\Delta t} \circ e^{A\Delta t/2}(u_n) = [e^{A\Delta t/2} \circ e^{B\Delta t} \circ e^{A\Delta t/2}]^n(u_0). \end{aligned} \tag{9}$$

 ISBN: 978-84-616-2723-3

We hope that both (8) and (9) converge towards the correct solution of (2), when the time step Δt tends to 0, that is,

$$u(t) = \lim_{\Delta t \to 0} [e^{A\Delta t} \circ e^{B\Delta t}]^n (u_0) = \lim_{\Delta t \to 0} [e^{A\Delta t/2} \circ e^{B\Delta t} \circ e^{A\Delta t/2}]^n (u_0). \tag{10}$$

Formally, Lie-Trotter splitting (8) converges as

$$\|u_n - u(t_n)\|_X \leq \mathcal{O}(\Delta t), \tag{11}$$

while the Strang splitting (9), converges as

$$\|u_n - u(t_n)\|_X \leq \mathcal{O}((\Delta t)^2). \tag{12}$$

The major task in what follows is to prove the convergence rates for the two operator splitting methods in Sobolev spaces. The main idea of the framework [11] is to use a standard argument from error estimation of numerical methods. We find an estimate of the *local error*, which is the error after performing one step with the operator splitting method, before we add up all the local errors from each step. This yield the global error, which is what we are after.

The keypoint in the new approach in [5] is to use error terms for numerical quadratures, to use the Peano kernel theorem for estimating the local errors in $H^s(\mathbb{R})$, where $H^s(\mathbb{R})$ is the Sobolev space where s is an arbitrary nonnegative integer. In addition, a Taylor series expansion and a variation of parameters formula are used to obtain the local estimates. These foundations yield an estimation of the local error which is delicate and elegant, and which involve the error forms in combination with differential calculus and estimation tools in $H^s(\mathbb{R})$.

We will investigate the Lie-Trotter splitting numerically for the given Burgers-Huxley equation. We will numerically check the convergence rates for the split step size Δt, in addition with other aspects for the numerical methods.

Applying the operator splitting method to (1), and splitting it into two subequations gives

$$v_t = A(v) = \epsilon v_{xx} \tag{13}$$

$$w_t = B(w) = \beta(1-w)(w-\gamma)w - \alpha w w_x \tag{14}$$

The analysis relies on a well-posedness theory for (1) in $H^s(\mathbb{R})$. For simplicity, we list the well-posedness requirements for (1) in addition with the assupmtions for u_0 and $u(t)$, ([1]),([4]).

Hypothesis 1 *(Local well-posedness). For a fixed time T, there exists $R > 0$such that for all u^0 in $H^k(\mathbb{R})$ with $\|u_0\|R$, there exists a unique strong solution u in $C([0,T], H^k)$ of (1). In addition, for the initial data u_0 there exists a constant $K(R,T) < \infty$, such that*

$$\|\tilde{u}(t) - u(t)\|_{H^k} \leq K(R,T)\|\tilde{u}_0 - u_0\|_{H^k}$$

 ISBN: 978-84-616-2723-3

for two arbitrary solutions u and $\tilde{u}$, corresponding to two different initial data $\tilde{u}_0$ and u_0.

The requirement in (15) is the same as requering that u_0 is local Lipschitz continuous. The last hypothesis requires that the solution and the initial data are bounded in the Sobolev spaces.

Hypothesis 2 *(Boundedness). The solution $u(t)$ and the initial data u_0 of (1) are both in $H^k(\mathbb{R})$, and are bounded as*

$$\|u(t)\|_{H^k} \leq R < \rho \text{ and } \|u_o\|_{H^k} \leq C < \infty, \tag{15}$$

for $0 \leq t \leq T$.

We define the following set of integers, which we keep fixed throughout this section,

$$s \geq 1, \quad p = s + 2l - 1, \quad , q = p - l \tag{16}$$

where $l \geq 2$.

We will use the following theorem and lemmas to estimate the local error for the Lie-Trotter splitting for the Burgers-Huxley equation.

Theorem 1 *(Peano Kernel theorem) If f is in $C^{n+1}([a,b])$ and I is a quadrature rule that integrates all p in $\mathbb{P}_n$ exactly,then*

$$E(f) \quad = \quad I(f) - \int_a^b f(x)dx = \frac{1}{n!}\int_a^b f^{n+1}(t)K(t)dt. \tag{17}$$

where $K(t) = E_x((x-t)_+^n)$ is the Peano kernel.

Lemma 2 *If u is in $H^s(\mathbb{R})$ for $s \geq 1$, then u is in $L^\infty(\mathbb{R})$. Moreover,*

$$\|u\|_{L^\infty} \leq \frac{1}{\sqrt{2}}\|u\|_{H^1} \leq C_s\|u\|_{H^s}, \tag{18}$$

where C_s depends only on s.

Lemma 3 *The space $H^s(\mathbb{R})$ is a Banach algebra for $s \geq 1$. In particular, if u, v are in $H^s(\mathbb{R})$ for $s \geq 1$, then*

$$\|uv\|_{H^s} \leq C_s\|u\|_{H^s}\|u\|_{H^s},$$

where where C_s depends only on s.

 ISBN: 978-84-616-2723-3

4 Regularity results for Burgers-Huxley Equation

We will present and prove several results to estimate the local error for the Lie-Trotter splitting for the Burgers-Huxley equation. We need to show that there exists a small time step Δt for the solutions $e^{At}(v_0)$ and $e^{Bt}(w_0)$ in a Sobolev spaces.

4.1 Results for the Nonlinear Part

Lemma 5 *For p and q in (16) assume the solution $e^{Bt}(w_0) = w(t)$ of (14) with initial data w_0 in $H^p(\mathbb{R})$, satisfies $\|e^{Bt}(w_0)\|_{H^q} \leq \alpha$ for $0 \leq t \leq \Delta t$. Then $e^{Bt}(w_0)$ is in $H^p(\mathbb{R})$ and in particular*

$$\|e^{Bt}(w_0)\|_{H^q} \leq e^{c\beta t}\|w_0\|_{H^p}, \tag{19}$$

where β and c is independent of w_0 and Δt.

Lemma 6 *Assume $\|w_0\|_{H^k} \leq K$ for some $k \geq 1$. Then there exists $\bar{t}(K) > 0$ such that $\|e^{Bt}(w_0)\|_{H^k} \leq 2K$ for $0 \leq t \leq \bar{t}(K)$.*

Proof 1 *By doing the same calculations as in the proof of Lemma (5) with k instead of p and using the bound for u_0 in $H^k(\mathbb{R})$, we arrive with the following inequality*

$$\|w(t)\|_{H^k}\frac{d}{dt}\|w(t)\|_{H^k} \leq c\|w(t)\|^4_{H^k}, \tag{20}$$

which simplifies to

$$\frac{d}{dt}\|w(t)\|_{H^k} \leq c\|w(t)\|^3_{H^k}. \tag{21}$$

By comparing with the solution of the differential equation $y' = cy^3$, we see that if we want $\|e^{Bt}(w_0)\|_{H^k} \leq 2K$, we must integrate the above inequality a time $\bar{t}$ which is dependent on the bound K.

To prove the convergence rates of the Lie-Trotter splitting, we need to expand $e^{Bt}(w_0)$ using the Taylor series expansion. Thus, $e^{Bt}(w_0)$ needs to be continuous, such that the expansions are valid. The following lemma proves the sufficient continuity.

Lemma 7 *If $\|w_0\|_{H^{s+2}} \leq C$ for $s \geq 1$, then there exists $\bar{t}$ depending on C, such that the solution $w(t)$ of the (14) is $C^2([0,\bar{t}], H^s)$.*

Proof 2 *Let t be in $[0,\bar{t}]$, with $\bar{t}$ from Lemma (6), and define*

$$\tilde{w}(t) = w_0 + tB(w_0) + \int_0^t (t-s)dB(w(s))[B(w(s))]ds, \tag{22}$$

 ISBN: 978-84-616-2723-3

where $dB(.)[.]$ is the Fréchet derivative. Calculating the second derivative of $\tilde{w}$, gives

$$\begin{aligned} \tilde{w}_{tt} &= dB(w(s))[B(w(s))] \\ &= -3\beta^2 B(w) + 2\beta(1+\gamma)wB(w) - \beta\gamma B(w) - \alpha w B(w)_x - \alpha B(w) w_x \end{aligned} \tag{23}$$

from which we have that $\tilde{w}$ is in $C^2([0,\bar{t}], H^s)$. To prove that $w = \tilde{w}$, we must show that the two functions satisfies the same differential equation and the same initial conditions. By differentiation (14) with respect to t, we get

$$\begin{aligned} w_{tt} &= B(w)_t = (-\beta w^3 + \beta(1+\gamma)w^2 - \beta\gamma w - \alpha w w_x)_t \\ &= -3\beta w^2 w_t + 2\beta(1+\gamma) w w_t - \beta\gamma w_t - \alpha w_t w_x - \alpha w w_{xt} \\ &= \tilde{w}_{tt}, \end{aligned}$$

which shows that w and $\tilde{w}$ satisfies the same equation. From the definition of $\tilde{w}$, we see that $\tilde{w}(0) = u_0$ and $\tilde{w}_t(0) = B(u_0) = w_t$. Thus we have shown that $w = \tilde{w}$.

8 Stability in H^s space

Lemma 9 *Let u_1, $\tilde{u}_1$ be the Lie-Trotter splitting solution with initial data u_0, $\tilde{u}_0$ in H^s. Then*

$$\|u_1 - \tilde{u}_1\|_{H^s} \leq e^{L\Delta t}\|u_0 - \tilde{u}_0\|_{H^s}, \tag{24}$$

where $L = K\max\{\|u_1\|_{H^s}, \|\tilde{u}_1\|_{H^s}\}$

Proof 3 *Since the linear part preserves the H^s norm, we only need to compare nonlinearities in Lie-Trotter splitting solutions. The nonlinear term has Lipshitz constant L which is bounden by Lemma (6). Finally, Gronwall's Lemma implies the bound in (24).*

10 Local error in H^s space

Lemma 11 *Let $s \geq 1$ be an integer and hypothesis 2 holds for $k = s+2$ for the solution $u(t) = e^{(A+B)\Delta t}(u_0)$ of (1). If the initial data u_0 is in $H^{s+2}(\mathbb{R})$, then the local error of the Lie-Trotter splitting (8) is bounded in $H^s(\mathbb{R})$ by*

$$\|e^{A\Delta t}(e^{B\Delta t}(u_0)) - e^{(A+B)\Delta t}(u_0)\|_{H^s} \leq C\Delta t^2, \tag{25}$$

where C only depends on $\|u_0\|_{H^{s+2}}$.

 ISBN: 978-84-616-2723-3

12 Global error in H^s space

Theorem 2 *Suppose that the exact solution $u(\cdot,t)$ of Equation (1) is in H^{s+2} for $0 \leq t \leq T$. Then Lie-Trotter splitting solution u_n has first order global error for $\Delta t < \bar{\Delta} t$ and $t_n = n\Delta t \leq T$,*

$$\|u_n - u(\cdot, t_n)\|_{H^s} \leq G\Delta t, \tag{26}$$

where G only depends on $\|u_0\|_{H^{s+2}}$ and T.

13 Numerical results

By applying the Lie-Trotter splitting to Burgers-Huxley equation, we obtain the two subequations

$$\begin{aligned} v_t &= A(v) = \epsilon v_{xx} \\ w_t &= B(w) = \beta(1-w)(w-\gamma)w - \alpha w w_x \end{aligned}$$

which are solved subequently for small time steps Δt.

We will use the Chebyshev Differentiation Matrices for the first and the second derivative of u in (13) and (14). For the second part (nonlinear part), we apply the semi-implicit RK scheme.

We consider the Burgers-Huxley equation with $\alpha = \beta = 1$, $\gamma = 0.5$ and initial and boundary conditions in the following form [16]

$$\begin{aligned} u(x,0) &= sin(\pi x), \;\; 0 \leq x \leq 1 \\ u(0,t) &= 0, \;\; 0 \leq t \leq T. \end{aligned} \tag{27}$$

The time step length $\Delta t = 0.001$ is used for the numerical experiment. The Figure 1 and Figure 2 show the layer behaviour of the problem at different values of time t and ϵ.

 ISBN: 978-84-616-2723-3

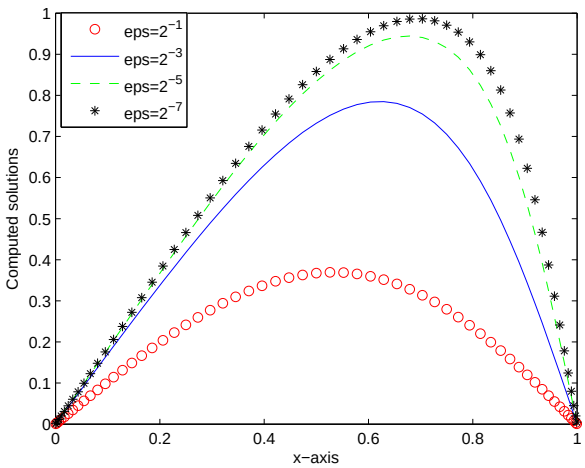


Figure 1: Computed solutions of Burgers-Huxley equation for different values of ϵ at T= 0.2.

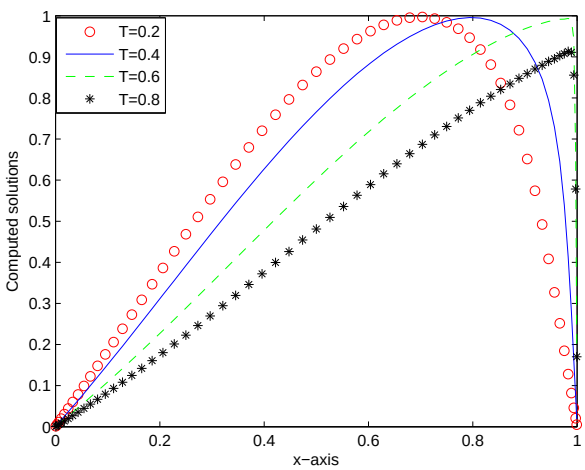


Figure 2: Computed solutions of Burgers-Huxley equation for different values of time at $\epsilon = 2^{-9}$.

 ISBN: 978-84-616-2723-3

References

[1] C. A. Ambrosetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge UP, Cambridge, 1995.

[2] H.Holden,C. Lubich and N. H.Risebro , *Operator splitting for partial differential equations with Burger nonlinerity*, Mathematics of Computation. **82** (2013) 173–185.

[3] H.Holden,C. Lubich , N. H.Risebro and T.Tao, *Operator splitting for the KDV equation*, Mathematics of Computation. **80** (2011) 821–846.

[4] A. Kolmogorov, I. Petrovskii and N. Piskunov, *Moscow Univ. Bull. Math.*, Moscow Univ. Bull. Math, 1937

[5] E. B. Nilsen, *On Operator Splitting for the Viscous Burgers' and the Korteweg-de Vries Equations*, Master of Science in Physics and Mathematics. (2011)

[6] G. I. Marchuk, *Methods of splitting*, Nauka, Moscow, 1988.

[7] G. Strang, *On the construction and comparison of different splitting schemes*, SIAM J. Numer. Anal. **5 (3)** (1968) 506–517.

[8] M. Mimura, T. Nakaki and K. Tomeada , *A numerical approach to interface curves for some nonlinear diffusion equations*, Japan. J. Appl. MAth. **1** (1968) 93–139.

[9] R. S. Marinova, C. I. Christov and T. T. Marinov , *A fully coupled solver for incompressible Navier-Stokes equations using operator splitting*, Int. J. Comput. Fluid Dyn. **17 (5)** (2003) 71–385.

[10] K. Hvistendahl Karlsen, A. Lie, H. F. Nordhaug and H. K. Dahle , *Operator splitting methods for systems of convection-diffusion equations: Nonlinear error mechanisms and correction strategies*, J. Comput. Phys. **173** (2001) 636–663.

[11] C. I. Christov, and R. S. Marinova , *Implicit vectorial operator splitting for incompressible Navier-Stokes equations in primitive variables*, J. Comput. Technol. **6 (4)** (2001) 92–119.

[12] J. G. Verwer, and B. Sportisse, *A note on operator splitting in a stiff linear case*, CWI, Amsterdam, Netherlands, MAS-R9830 1998.

[13] K. A. Bagrinovski, and S. K. Godunov , *Difference schemesfor multidimensional problems*, Dokl.Akad.Nauk SSSR(NS). **115** (1957) 413–433.

[14] G. I. Marchuk, *Methods of splitting*, Nauka, Moscow, 1998.

 ISBN: 978-84-616-2723-3

[15] G. Strang , *On the construction and comparison of different splitting schemes*, SIAM J. Numer. Anal. **5** (1968) 506–517.

[16] Ram Jiwari and R. C. Mittal , *A Higher Order Numerical Scheme for Singularly Perturbed Burger-Huxley Equation*, J. Appl. Math. and Informatics Vol. **29** (2011) 813–829.

 ISBN: 978-84-616-2723-3

Proceedings of the 14th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2014
3–7 July, 2014.

A high order in space uniformly convergent method for parabolic singularly perturbed reaction-diffusion systems

C. Clavero[1] and J.L. Gracia[1]

[1] *IUMA and Department of Applied Mathematics, University of Zaragoza, Zaragoza, Spain*

emails: clavero@unizar.es, jlgracia@unizar.es

Abstract

In this work we approximate the solution of a two-point boundary value singularly perturbed system with two parabolic equations of reaction-diffusion type, coupled in the reaction term. This class of problems typically exhibits two overlapping boundary layers at both end points of the spatial domain. The numerical scheme combines the backward Euler method to discretize in time and a hybrid finite difference scheme, defined on a special nonuniform mesh, to discretize in space. The hybrid scheme uses two finite difference operators which yield a full discrete monotone scheme. A truncation error argument is used to proved that the numerical method is uniformly convergent in the discrete maximum norm, having first and third order of convergence in time and space, respectively. A test example is showed, which illustrates the order of convergence of the numerical method.

Key words: singular perturbation, parabolic reaction-diffusion systems, special nonuniform mesh uniform convergence, high order method
MSC 2000: 65M06, 65N06, 65N12

1 Introduction

In this paper we consider 1D parabolic singularly perturbed boundary value problems of type

$$\begin{cases} \mathcal{L}_{\boldsymbol{\varepsilon}}\mathbf{u} \equiv \dfrac{\partial \mathbf{u}}{\partial t} + \mathcal{L}_{x,\boldsymbol{\varepsilon}}\mathbf{u} = \mathbf{f}, & (x,t) \in G = \Omega \times (0,T] \equiv (0,1) \times (0,T], \\ \mathbf{u}(0,t) = \mathbf{0}, \ \mathbf{u}(1,t) = \mathbf{0}, & \forall t \in [0,T], \\ \mathbf{u}(x,0) = \mathbf{0}, & \forall x \in \Omega, \end{cases} \tag{1}$$

 ISBN: 978-84-616-2723-3

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)^T$ is the singular perturbation parameter with $0 < \varepsilon_1 \leq \varepsilon_2 \leq 1$, $\mathbf{f}(x,t) = (f_1(x,t), f_2(x,t))^T$, and the differential operator $\mathcal{L}_{x,\boldsymbol{\varepsilon}}$ is defined by

$$\mathcal{L}_{x,\boldsymbol{\varepsilon}} \equiv \begin{pmatrix} -\varepsilon_1 \frac{\partial^2}{\partial x^2} & \\ & -\varepsilon_2 \frac{\partial^2}{\partial x^2} \end{pmatrix} + A, \qquad A = \begin{pmatrix} a_{11}(x,t) & a_{12}(x,t) \\ a_{21}(x,t) & a_{22}(x,t) \end{pmatrix}.$$

We assume that the data problem are smooth functions and enough compatibility are satisfied to guarantee the regularity of the solution of (1). In our analysis we need that the continuous and discrete problems satisfy a comparison principle and then we also assume that

$$\begin{aligned} &a_{i1} + a_{i2} \geq 0, \quad a_{ii} > 0, \quad i = 1, 2, \\ &a_{ij} \leq 0 \ \text{ if } \ i \neq j. \end{aligned}$$

The exact solution $\mathbf{u}$ has two overlapping boundary layers at $x = 0$ and $x = 1$ of width $\mathcal{O}(\sqrt{\varepsilon_i}\ln(1/\varepsilon_i))$, $i = 1, 2$ (see [5, 9]). Then, to obtain good approximations for any value of the diffusion parameter $\boldsymbol{\varepsilon}$, uniformly convergent methods (see [5, 7, 8, 9] and references therein) are necessary.

In [5, 9] the backward Euler method and central differences on a Shishkin mesh are used to approximate problem (1) and it is proved that the numerical method is a first order uniformly convergent scheme. In the general case of different diffusion parameters, up to our knowledge, central differences has been only used to approximate problem (1). Here we are interested into constructing a numerical method on a special mesh to solve (1), giving higher order convergence for the space variable than this one associated to the classical central difference scheme. In the steady version of problem (1) high order schemes have been designed and analyzed in the literature. We can refer to [2], where an almost third order uniformly convergent finite difference scheme, on a piecewise uniform Shishkin mesh, was designed to solve a reaction-diffusion coupled system when $\varepsilon_1 = \varepsilon_2$, which was extended in [3] to the general case $\varepsilon_1 \neq \varepsilon_2$. Some of the finite difference operators of those paper are used in this work.

Henceforth, C denotes a generic positive constant independent of $\boldsymbol{\varepsilon}$ and also of the discretization parameters N and M. We only use the (discrete) maximum norm $\|f\|_D = \max_{x\in D} |f(x)|$, $\|\mathbf{f}\|_D = \max\{\|f_1\|_D\,, \|f_2\|_D\}$ with $\mathbf{f} = (f_1, f_2)^T$.

2 Uniform convergence of the numerical scheme

Before analyzing the convergence of the numerical scheme, it is necessary to dispose of appropriate bounds of the derivatives of the solution of (1). The asymptotic behavior of the solution, with respect to the diffusion parameters, is related to the exponential boundary layer functions

$$B_{\varepsilon_i}(x) = e^{-x/\sqrt{\varepsilon_i}} + e^{-(1-x)/\sqrt{\varepsilon_i}}, \quad i = 1, 2.$$

 ISBN: 978-84-616-2723-3

In [1, 5] it is proved that

$$\begin{aligned}
&|u_i^{(0,k_0)}(x,t)| \le C(1+B_{\varepsilon_2}(x)), \quad i=1,2,\ 0\le k_0\le 3,\\
&|u_1^{(k,0)}(x,t)| \le C(1+\varepsilon_1^{-k/2}B_{\varepsilon_1}(x)+\varepsilon_2^{-k/2}B_{\varepsilon_2}(x)),\ 1\le k\le 6\\
&|u_2^{(k,0)}(x,t)| \le C(1+\varepsilon_2^{-k/2}B_{\varepsilon_2}(x)),\ k=1,2,\\
&|u_2^{(k,0)}(x,t)| \le C(1+\varepsilon_2^{-1}(\varepsilon_1^{(2-k)/2}B_{\varepsilon_1}(x)+\varepsilon_2^{(2-k)/2}B_{\varepsilon_2}(x))),\ \ 3\le k\le 6.
\end{aligned} \tag{2}$$

To define the numerical scheme, the first step is to construct the mesh, which is denoted by $\bar{G}^{N,M}=\bar{\Omega}^N\times\bar{\omega}^M$, where $\bar{\omega}^M=\{t_k=k\tau,\ 0\le k\le M,\ \tau=T/M\}$ and M is a positive integer. Here a modified Shishkin mesh (see [6, 10]) is defined, which uses two transition parameters given by

$$\sigma_2=\min\{1/4,4\sqrt{\varepsilon_2}\ln N\},\quad \sigma_1=\min\{\sigma_2/2,4\sqrt{\varepsilon_1}\ln N\}, \tag{3}$$

and the grid condenses in the layer regions.

Using a suitable generating function $\aleph$, the grid points are given by $x_j=\aleph(j/N),\ j=0,1,\ldots,N/2$, with $\aleph\in\mathcal{C}^1[0,1/2]$ and $N=8k$, with k a positive integer.

We first consider the case when $\sigma_2\neq 1/4$. Then, we extend the definition of the Vulanović-Shishkin type mesh for the case of a two point boundary value problem with a single equation (see [6, 10]), by using

$$\aleph(z)=\begin{cases} 8z\sigma_1, & \text{if } z\in[0,1/8],\\ p_1(z-1/8)^3+8\sigma_1(z-1/8)+\sigma_1, & \text{if } z\in[1/8,1/4],\\ p_2(z-1/4)^3+p_3(z-1/4)+\sigma_2, & \text{if } z\in[1/4,1/2],\end{cases} \tag{4}$$

where the value of p_1,p_2 and p_3 are calculated by imposing that $\aleph(1/4)=\sigma_2$, $\aleph(1/2)=1/2$ and the mesh is symmetric with respect to the mesh point $1/2$.

Imposing that $\aleph(z)$ be an increasing function, that $h_j\ge h_{j-1}$ for $j=2,\ldots,N/2$ (since the layer is at x=0), where $h_j=x_j-x_{j-1},\ j=1,2,\ldots,N$, it follows that

$$p_1=8^3(\sigma_2-2\sigma_1),\ p_2=64(1/2-7\sigma_2+10\sigma_1),\ p_3=24\sigma_2-40\sigma_1, \tag{5}$$

and it must be satisfied the condition

$$1/2-7\sigma_2+10\sigma_1\ge 0. \tag{6}$$

Then, it holds that $h_j\le CN^{-1}$ for $j=1,\ldots,N$ and

$$|h_{j+1}-h_j|\le\begin{cases} CN^{-2}\sqrt{\varepsilon_2}\ln N, & \text{for } j=N/8,\ldots,N/4-1,\\ CN^{-2}, & \text{for } j=N/4,\ldots,N/2-1.\end{cases} \tag{7}$$

 ISBN: 978-84-616-2723-3

Otherwise, we have that $1/2 < 1/2 + 10\sigma_1 < 7\sigma_2 \leq 28\sqrt{\varepsilon_2}\ln N$ and therefore $\varepsilon_2^{-1/2} \leq 56\ln N$. If either (6) does not hold or $\sigma_2 = 1/4$, we change the definition of the mesh, taking

$$\widetilde{\aleph}(z) = \begin{cases} 8z\sigma_1, & \text{if } z \in [0, 1/8], \\ 8^3(1/2 - 4\sigma_1)(z - 1/8)^3 + 8\sigma_1(z - 1/8) + \sigma_1, & \text{if } z \in [1/8, 1/2], \end{cases} \tag{8}$$

that satisfies the same conditions as before.

On the previous Vulanović-Shishkin meshes, we define a finite difference scheme. The values of the numerical solution for the initial time $t = 0$ and at $x = 0$ and $x = 1$ are defined by

$$\mathbf{U}(x_j, 0) = \mathbf{U}(0, t_n) = \mathbf{U}(1, t_n) = \mathbf{0}, \quad \text{for } j = 0, \ldots, N \text{ and } n = 0, \ldots, M.$$

At the interior points, we consider a hybrid finite difference operator $\mathbf{L}^{N,M} = (L_1^{N,M}, L_2^{N,M})$ defined by

$$L_i^{N,M}\mathbf{U}(x_j, t_n) = Q_i^{N,M}(f_i(x_j, t_n)), \quad i = 1, 2, \tag{9}$$

for $j = 1, \ldots, N-1$ and $n = 1, \ldots, M$, where

$$L_i^{N,M}\mathbf{U}(x_j, t_n) \equiv Q_i^{N,M}(D_t^- U_i(x_j, t_n)) + L_{x,i}^{N,M}\mathbf{U}(x_j, t_n),$$

D_t^- is the backward finite difference, $L_{x,i}^{N,M}$ is defined as

$$\begin{aligned} L_{x,i}^{N,M}\mathbf{U}(x_j, t_n) &= r_{i,j}^{-,n}U_i(x_{j-1}, t_n) + r_{i,j}^{c,n}U_i(x_j, t_n) + r_{i,j}^{+,n}U_i(x_{j+1}, t_n) \\ &\quad + Q_i^{N,M}(a_{i,3-i}(x_j, t_n)U_{3-i}(x_j, t_n)), \end{aligned} \tag{10}$$

and

$$Q_i^{N,M}(Z(x_j, t_n)) = q_{i,j}^1 Z(x_{j-1}, t_n) + q_{i,j}^2 Z(x_j, t_n) + q_{i,j}^3 Z(x_{j+1}, t_n), \ i = 1, 2. \tag{11}$$

The coefficients $r_{i,j}^{*,n}$ with $* = \{-, c, +\}$ of the scheme are given by

$$\begin{aligned} r_{i,j}^{+,n} &= -2\varepsilon_i/(h_{j+1}(h_j + h_{j+1})) + q_{i,j}^3 a_{i,i}(x_{j+1}, t_n), \\ r_{i,j}^{-,n} &= -2\varepsilon_i/(h_j(h_j + h_{j+1})) + q_{i,j}^1 a_{i,i}(x_{j-1}, t_n), \\ r_{i,j}^{c,n} &= q_{i,j}^1 a_{i,i}(x_{j-1}, t_n) + q_{i,j}^2 a_{i,i}(x_j, t_n) + q_{i,j}^3 a_{i,i}(x_{j+1}, t_n) - r_{i,j}^{-,n} - r_{i,j}^{+,n}, \end{aligned} \tag{12}$$

which depend on the value of $q_{i,j}^k$, $k = 1, 2, 3$. We use different values, depending on the equation and the ratio between the diffusion and the discretization parameters, in order that the matrix associated to the scheme be an M-matrix and hence the scheme is monotone. Concretely, we consider the following sets of values: CD $= \{q_{i,j}^1 = q_{i,j}^3 = 0, \ q_{i,j}^2 = 1\}$ corresponds to the standard central difference approximation; HS corresponds to the choice of a HODIE (High Order via Differential Identity Expansion) scheme [6], for which

$$\begin{aligned} q_{i,j}^1 &= \frac{1}{6}\left(1 - \frac{h_{j+1}^2}{h_j(h_j + h_{j+1})}\right), \\ q_{i,j}^3 &= \frac{1}{6}\left(1 - \frac{h_j^2}{h_{j+1}(h_j + h_{j+1})}\right), \\ q_{i,j}^2 &= 1 - q_{i,j}^1 - q_{i,j}^3. \end{aligned}$$

 ISBN: 978-84-616-2723-3

Using these two set of values, the hybrid scheme on the Vulanović-Shishkin mesh generated by $\aleph(z)$ is defined taking

$$\begin{aligned}
&\text{HS}, \quad \text{if } x_j \in (0,\sigma_1)\cup(1-\sigma_1,1), \text{ and } i=1,2,\\
&\text{CD}, \quad \text{if } x_j \in [\sigma_1,\sigma_2)\cup(1-\sigma_2,1-\sigma_1], \ h_{max}^2\left(\|a_{11}\|_{\bar G}+\frac{1}{\tau}\right)\geq 6\varepsilon_1, \text{ and } i=1,\\
&\text{HS}, \quad \text{if } x_j \in [\sigma_1,\sigma_2)\cup(1-\sigma_2,1-\sigma_1], \ h_{max}^2\left(\|a_{11}\|_{\bar G}+\frac{1}{\tau}\right)< 6\varepsilon_1, \text{ and } i=1,\\
&\text{HS}, \quad \text{if } x_j \in [\sigma_1,\sigma_2)\cup(1-\sigma_2,1-\sigma_1], \text{ and } i=2,\\
&\text{CD}, \quad \text{if } x_j \in [\sigma_2,1-\sigma_2], \ H_{max}^2\left(\|a_{11}\|_{\bar G}+\frac{1}{\tau}\right)\geq 6\varepsilon_1, \text{ and } i=1,\\
&\text{HS}, \quad \text{if } x_j \in [\sigma_2,1-\sigma_2], \ H_{max}^2\left(\|a_{11}\|_{\bar G}+\frac{1}{\tau}\right)< 6\varepsilon_1, \text{ and } i=1,\\
&\text{CD}, \quad \text{if } x_j \in [\sigma_2,1-\sigma_2], \ H_{max}^2\left(\|a_{22}\|_{\bar G}+\frac{1}{\tau}\right)\geq 6\varepsilon_2, \text{ and } i=2,\\
&\text{HS}, \quad \text{if } x_j \in [\sigma_2,1-\sigma_2], \ H_{max}^2\left(\|a_{22}\|_{\bar G}+\frac{1}{\tau}\right)< 6\varepsilon_2, \text{ and } i=2,
\end{aligned}$$

where $h_{max}=\max\limits_{N/8+1\leq j\leq N/4} h_j$, $\quad H_{max}=\max\limits_{N/4+1\leq j\leq N/2} h_j$. It is easy to prove that

$$r_{i,j}^{c,n}>0, \quad r_{i,j}^{-,n}\leq 0, \quad r_{i,j}^{+,n}\leq 0, \ r_{i,j}^{c}+r_{i,j}^{-}+r_{i,j}^{+}>0 \quad q_{i,j}^{k}\geq 0, \ k=1,2,3. \tag{13}$$

On the Vulanović-Shishkin mesh generated by $\widetilde{\aleph}(z)$, we consider the following operators

$$\begin{aligned}
&\text{HS}, \quad \text{if } \widetilde{x}_j \in (0,\sigma_1)\cup(1-\sigma_1,1), \text{ and } i=1,2,\\
&\text{CD}, \quad \text{if } \widetilde{x}_j \in [\sigma_1,1-\sigma_1], \ \widetilde{H}_{max}^2\left(\|a_{11}\|_{\bar G}+\frac{1}{\tau}\right)\geq 6\varepsilon_1, \text{ and } i=1,\\
&\text{HS}, \quad \text{if } \widetilde{x}_j \in [\sigma_1,1-\sigma_1], \ \widetilde{H}_{max}^2\left(\|a_{11}\|_{\bar G}+\frac{1}{\tau}\right)< 6\varepsilon_1, \text{ and } i=1,\\
&\text{HS}, \quad \text{if } \widetilde{x}_j \in [\sigma_1,1-\sigma_1], \text{ and } i=2,
\end{aligned}$$

where $\widetilde{H}_{max}=\max\limits_{N/8+1\leq j\leq N/2} \widetilde{h}_j$. In this case, if $M\leq CN^2/\ln^2 N$, then the set HS suffices so that (13) is true for the coefficients of the second equation in the region $[\sigma_1,1-\sigma_1]$.

In [1] the following Theorem giving the uniform convergence of the hybrid scheme is proved. The proof is based on a stability and a truncation error argument, that uses the estimates of the derivatives given in (2).

Theorem 1. *Assume that $M\leq CN^2/\ln^2 N$. Let $\mathbf{u}$ be the solution of problem (1) and $\mathbf{U}$ the solution of the monotone numerical method (9) on the mesh (4) when (6) holds and on the mesh (8) in other case. Then, the error at the grid points satisfies*

$$\|\mathbf{U}-\mathbf{u}\|_{\bar G^{N,M}}\leq C(M^{-1}+CN^{-2}\min\{\varepsilon_2,N^{-2}M\}+N^{-4}\ln^4 N+MN^{-5}\ln^3 N).$$

Remark 2. *Note that if $M\leq CN$, the scheme is a first order in time and almost third order in space uniformly convergent scheme.*

 ISBN: 978-84-616-2723-3

3 Numerical results

In this section, we show the numerical results obtained for a test problem. The initial and boundary conditions are zero, the reaction matrix is given by

$$A = \begin{pmatrix} 2(x^2+1)+\sin(\pi x) & \cos(\pi x)-2 \\ -4x & e^{x+1} \end{pmatrix},$$

and the right-hand side is $f(x,t) = (x(1-x), tx)^T$.

In all tables we show the results for two pairs of values of the diffusion parameters ε_1 and ε_2. In the first case, $\varepsilon_1 = 2^{-20}$ and $\varepsilon_2 = 2^{-14}$; then $\sigma_2 < 1/4, \sigma_1 < \sigma_2/2$ and there exist two overlapping boundary layer, but the restriction (6) does not hold. In the second one, $\varepsilon_1 = 2^{-22}$ and $\varepsilon_2 = 2^{-17}$ and therefore $\sigma_2 < 1/4, \sigma_1 < \sigma_2/2$, again there exist two overlapping boundary layer, but the restriction (6) holds.

The exact solution is unknown and we use a variant of the two-mesh principle (see [4] for a justification of this method) to approximate the maximum pointwise errors. Then, we calculate $\{\hat{\mathbf{U}}_j^n\}$, the numerical solution on the mesh $\{(\hat{x}_j, \hat{t}_n)\}$ containing the original mesh points and its midpoints, i.e.,

$$\begin{array}{llll} \hat{x}_{2j} = x_j, & j = 0,\ldots,N, & \hat{x}_{2j+1} = (x_j + x_{j+1})/2, & j = 0,\ldots,N-1, \\ \hat{t}_{2n} = t_n, & n = 0,\ldots,M, & \hat{t}_{2n+1} = (t_n + t_{n+1})/2, & n = 0,\ldots,M-1. \end{array}$$

The maximum errors at the mesh points of the coarse mesh are approximated by computing the following two-mesh differences

$$d_{i,N,M} = \max_{0\le n\le M} \max_{0\le j\le N} |U_{i,j}^n - \hat{U}_{i,2j}^{2n}|, \quad i = 1,2,$$

and the orders of convergence are calculated by

$$q_i = \frac{\log(d_{i,N,M}/d_{i,2N,2M})}{\log 2}, \quad i = 1,2. \tag{14}$$

Table 1 displays the maximum two-mesh differences doubling the values of N and M; from it we deduce the first order of convergence in agreement with Theorem 1.

To see the influence of errors associated to the space discretization on the global error, we calculate the following orders of convergence

$$q_i^* = \frac{\log(d_{i,N,M}/d_{i,2N,4M})}{\log 2}, \quad q_i^{**} = \frac{\log(d_{i,N,M}/d_{i,2N,8M})}{\log 2}, \quad i = 1,2. \tag{15}$$

Tables 2 and 3 display the maximum two-mesh differences in these cases; from them, we observe second and third order of convergence, respectively, according with the ratio of the time step size. These results confirm the high order approximation in space of the numerical scheme (9), according with Theorem 1.

 ISBN: 978-84-616-2723-3

Table 1: Maximum two-mesh differences and orders of convergence

		N=32 M=8	N=64 M=16	N=128 M=32	N=256 M=64	N=512 M=128	N=1024 M=256
$\varepsilon_1 = 2^{-20}$	$d_{1,N,M}$	1.951E-3	1.046E-3	5.434E-4	2.771E-4	1.399E-4	7.030E-5
	q_1	0.899	0.944	0.972	0.986	0.993	
$\varepsilon_2 = 2^{-14}$	$d_{2,N,M}$	1.626E-3	8.688E-4	4.672E-4	2.468E-4	1.271E-4	6.456E-5
	q_2	0.904	0.895	0.921	0.957	0.978	
$\varepsilon_1 = 2^{-22}$	$d_{1,N,M}$	2.034E-3	1.086E-3	5.638E-4	2.873E-4	1.451E-4	7.288E-5
	q_1	0.905	0.946	0.973	0.986	0.993	
$\varepsilon_2 = 2^{-17}$	$d_{2,N,M}$	1.660E-3	8.807E-4	4.672E-4	2.468E-4	1.272E-4	6.456E-5
	q_2	0.914	0.915	0.921	0.957	0.978	

Table 2: Maximum two-mesh differences and orders of convergence

		N=32 M=8	N=64 M=32	N=128 M=128	N=256 M=512	N=512 M=2048	N=1024 M=8192
$\varepsilon_1 = 2^{-20}$	$d_{1,N,M}$	1.951E-3	5.429E-4	1.399E-4	3.524E-5	8.827E-6	2.208E-6
	q_1^*	1.845	1.956	1.989	1.997	1.999	
$\varepsilon_2 = 2^{-14}$	$d_{2,N,M}$	1.626E-3	4.672E-4	1.271E-4	3.253E-5	8.182E-6	2.049E-6
	q_2^*	1.799	1.877	1.966	1.991	1.998	
$\varepsilon_1 = 2^{-22}$	$d_{1,N,M}$	2.034E-3	5.641E-4	1.451E-4	3.654E-5	9.150E-6	2.289E-6
	q_1^*	1.851	1.959	1.990	1.997	1.999	
$\varepsilon_2 = 2^{-17}$	$d_{2,N,M}$	1.660E-3	4.671E-4	1.271E-4	3.254E-5	8.183E-6	2.049E-6
	q_2^*	1.829	1.877	1.966	1.991	1.998	

Acknowledgements

This research was partially supported by the project MEC/FEDER MTM 2010-16917 and the Diputación General de Aragón.

References

[1] C. Clavero, J.L. Gracia, *An improved uniformly convergent scheme in space for 1D parabolic reaction-diffusion systems*, submitted to publication.

[2] C. Clavero, J.L. Gracia, F.J. Lisbona, *High order schemes for reaction-diffusion singularly perturbed systems*, Lecture Notes in Comput. Sci. Engineering **69** (2009) 107–115.

[3] C. Clavero, J.L. Gracia, F.J. Lisbona, *An almost thrid order finite difference scheme for singularly perturbed reaction-diffusion systems*, J. Comp. Appl. Math. **234** (2010) 2501–2515.

Table 3: Maximum two-mesh differences and orders of convergence

		N=32 M=8	N=64 M=64	N=128 M=512	N=256 M=4096	N=512 M=32768	N=1024 M=262144
$\varepsilon_1 = 2^{-20}$	$d_{1,N,M}$	1.951E-3	2.767E-4	3.527E-5	6.180E-6	7.398E-7	7.123E-8
	q_1^{**}	2.818	2.972	2.513	3.062	3.376	
$\varepsilon_2 = 2^{-14}$	$d_{2,N,M}$	1.626E-3	2.467E-4	3.253E-5	4.095E-6	5.124E-7	6.405E-8
	q_2^{**}	2.721	2.923	2.990	2.999	3.000	
$\varepsilon_1 = 2^{-22}$	$d_{1,N,M}$	2.034E-3	2.878E-4	5.398E-5	1.260E-5	1.512E-6	1.462E-7
	q_1^{**}	2.821	2.415	2.099	3.059	3.370	
$\varepsilon_2 = 2^{-17}$	$d_{2,N,M}$	1.660E-3	2.467E-4	3.253E-5	4.095E-6	5.124E-7	6.406E-8
	q_2^{**}	2.750	2.923	2.990	2.999	3.000	

[4] P.A. Farrell, A. Hegarty, *On the determination of the order of uniform convergence*, in Proceedings of IMACS'91 **v. 2** (1991) 501–502.

[5] J.L. Gracia, F. Lisbona, *A uniformly convergent numerical scheme for a system of reaction-diffusion equations*, J. Comp. Appl. Math. **206** (2007) 1–16.

[6] D. Herceg, *Uniform fourth order difference scheme for a singular perturbation problem*, Numer. Math. **56** (1990) 675–693.

[7] N. Madden, M. Stynes, *A uniformly convergent numerical method for a coupled system of two singularly perturbed linear reaction-diffusion problems*, IMA J. Numer. Anal. **23** (2003) 627–644.

[8] 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.

[9] G.I. Shishkin, *Mesh approximation of singularly perturbed boundary-value problems for systems of elliptic and parabolic equations*, Comp. Maths. Math. Phys. **35** (1995) 429–446.

[10] R. Vulanović, *An almost sixth-order finite-difference method for semilinear singular perturbation problems*, Comp. Meth. Appl. Math. **4** (2004) 368–383.

Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2014 3–7July, 2014.

Energy-Efficient Allocation of Computing Node Slots in HPC Clusters through Evolutionary Multi-Criteria Decision Making

Alberto Cocaña-Fernández[1], José Ranilla[1] and Luciano Sánchez[1]

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

emails: cocanaalberto@gmail.com, ranilla@uniovi.es, luciano@uniovi.es

Abstract

Decision-making mechanisms for on-line allocation of computer node slots in HPC clusters are commonly based on simple knowledge-based systems comprised of individual sets of if-then rules. In contrast with previous works where these rules were designed using expert knowledge, an evolutionary learning algorithm is introduced in this paper that discovers the most appropriate knowledge base for a given load scenario. The proposed approach optimizes the quality of service and the number of node reconfigurations along with the energy consumption. An experimental study has been made using actual workloads from the Scientific Modelling Cluster at Oviedo University, and statistical evidence was found supporting the adoption of the new learning system.

Key words: Energy-efficient cluster computing; Multi-criteria decision making; Evolutionary algorithms

1 Introduction

High Performance Computing clusters have become a very important element in both scientific and industrial communities because they are an excellent platform for solving a wide range of problems through parallel and distributed applications [5]. Nowadays, HPC clusters are, in fact, the main architecture for supercomputers (as shown in Top500 architecture distribution[1]) due to the high performance of commodity microprocessors and networks, to the standard tools for high performance distributed computing, and to the lower price/performance ratio [39].

[1]November 2013 — TOP500 Supercomputer Sites, http://www.top500.org/lists/2013/11/

Nevertheless, this high performance comes at the price of consuming large amounts of energy. According to the U.S. Environmental Protection Agency [35], the consumption of data centers in USA was estimated at 61 billion kilowatt-hours (kWh) in 2006 for a total electricity cost of about $4.5 billion.

Large energy consumptions combined with notably increasing electricity prices in both EU [15] and USA [12] also have an important economical impact for IT companies, driving up power and cooling costs and forcing them to reduce operation costs [11, 34].

The environmental impact of the high energy consumption is also very significant. The EPA 2011 projected CO_2 emissions were 67.9 million metric tons [35]. Gartner estimates that the ICT industry accounts for 2 percent of global CO_2 emissions, a figure equivalent to aviation [19].

This environmental and economical impact is the main bottleneck constraining the expansion of supercomputing and data centers and, therefore, a powerful motivation to maximize the efficiency of clusters. Moreover, a side effect of reducing the energy consumption of clusters is the reduction in heat dissipation, what can increase reliability. Also, it produces a cascade effect reducing the consumption of auxiliary devices such as Power Supply Units, power distribution, cooling, lighting and building switchgear, what further encourages to look for energy efficiency in cluster computing [14].

Many methods have been proposed within the field of energy-efficient cluster computing following both static and dynamic approaches. An example of static approach is the development of low-power CPUs such as the IBM PowerPC A2 of IBM Blue Gene/Q [21, 25], or the use of GPUs and Intel Xeon Phi coprocessors. On the other hand, dynamic approaches adapt the cluster to its resource requirements at every given moment, thus saving energy when not needed [36]. An example is the Dynamic Voltage and Frequency Scaling (DVFS) technique, which reduces CPU voltage and frequency when the CPU is idle or under-used. This technique was used in [23, 22, 17, 29, 6, 20, 24, 7]. Other examples are the software frameworks to develop energy-efficient applications, such as [1, 32, 16, 28, 37], energy-efficient job schedulers [41, 40] and thermal-aware methods [3, 33].

However, the most relevant technique for this paper is the adaptive resource cluster, which consists mainly in switching on and off cluster compute nodes, adapting to the requested resources at every moment and, therefore, saving energy. This technique was first introduced in [31] for Load-Balancing clusters, and was also used in [8, 13, 4, 27, 18, 30] and in VMware vSphere[2] and Citrix XenServer hypervisors [3].

Recently it has also been applied to HPC clusters in [2, 10] or [38]. In these works, the decision-making mechanism for determining the adequate resources (e.g. number of compute node slots) at every moment is based on a simple Knowledge-based System (KBS)

[2]VMware Distributed Power Management Concepts and Use,
http://www.vmware.com/files/pdf/Distributed-Power-Management-vSphere.pdf

[3]Citrix XenServer - Efficient Server Virtualization Software,
http://www.citrix.com/products/xenserver/overview.html

 ISBN: 978-84-616-9216-3

comprised of an individual set of if-then rules. The KBS constantly monitors requested, idle and available resources. The rule base governing this system is made to depend on certain configuration parameters such as the time of inactivity to shutdown nodes. These parameters are tuned by hand, according to the experience of the administrator.

According to our own experience, these systems are not location-agnostic. In order to obtain the best energy saving, both the set of rules defining the system and the parameters on which the rules depend must be optimized for the actual load scenario.

Otherwise, the results would either interfere with the desired operation of the cluster or would not save as much energy as it could be possible. Because of this, we propose a cluster management system, that works with both OGE/SGE and PBS/TORQUE Resource Management Systems (RMS), whose decision-making mechanism shares the same rule set proposed in [10], as we consider it the soundest, but whose numerical parameters are obtained by means of a multiobjective evolutionary algorithm in a machine learning approach. The purpose of the learning is to fine-tune the KBS to the expected cluster activity, while complying with the preferences of the administrator in all QoS, energy saved and node reconfigurations.

The remainder of the paper is as follows. Section 2 explains the architecture of the solution proposed. Section 3 explains the learning algorithm used. Section 4 shows the experimental results. Section 5 concludes the paper and discusses the future work.

2 Architecture

The solution proposed consists on a service and an administration dashboard, coupled with a Database Management System, and deployed over an HPC cluster running a Resource Management System such as OGE/SGE or PBS/TORQUE. The underlying architecture of these clusters combines a master node and several computing nodes. Cluster users access the master node through a remote connection such as SSH and they submit jobs to the RMS. The RMS schedules jobs execution and when dispatched, jobs are assigned slots among the compute nodes, which are the ones actually running the job. Each slot represents a resource in the cluster, and depending on the RMS configuration the size of the resource ranges from a single CPU core to an entire host.

Figure 1 provides a high-level overview of the system components. The mission of the EEClusterd service is to periodically synchronize with the system status using various components and applications, and then use the Knowledge-based System to make decisions on whether any reconfiguration of the compute nodes must be performed. The administration dashboard is a Web application that displays current cluster status (nodes, queues, jobs, users), statistics, charts... and also allows the cluster administrator to switch on/off nodes manually and configure the system.

 ISBN: 978-84-616-9216-3

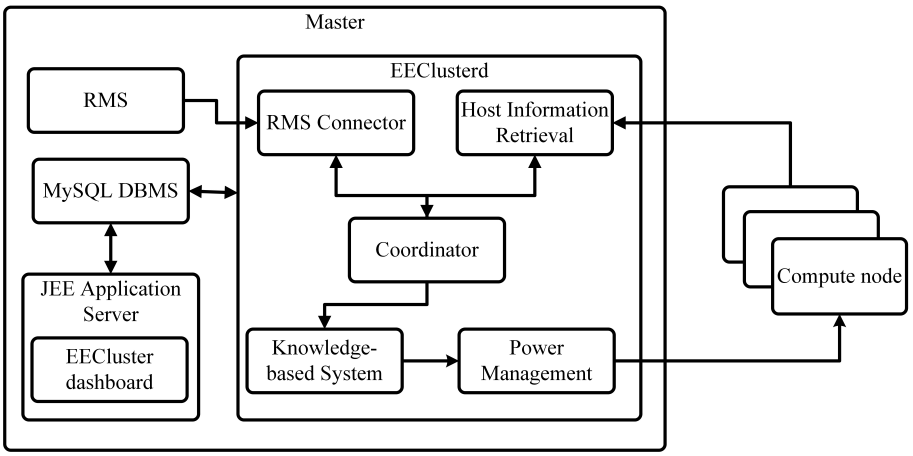


Figure 1: System components overview

2.1 Synchronization

The synchronization task of the service collects and keeps updated records of the RMS and of every compute node. RMS data includes the cluster parallel environments (OGE/SGE), queues, hosts, users, and completed, queued and running jobs. The service retrieves this information through the RMS connector, which uses multiple command line applications including *qhost* (hosts data in OGE/SGE), *pbsnodes* (hosts data in PBS/TORQUE), *qconf* (queues and parallel environments), *qacct* (users and jobs), *qstat* (current running and queued jobs), and also the PBS/TORQUE accounting records for completed jobs and the */etc/passwd* file for user data. Regarding hosts, the Host Information Retrieval module collects data of CPUs (*/proc/cpuinfo*), memory (*/proc/meminfo*), GPUs(through NVIDIA System Management Interface), Intel MIC devices (*micinfo*), PSUs power consumptions (through IPMI cards), and also the MAC address.

2.2 Power Management

The Power Management module is the responsible for switching on/off the nodes appointed by the Knowledge-based System. This can be done either using Ethernet cards or IPMI cards (Intelligent Platform Management Interface). With Ethernet cards, the power on order is carried out by sending the Ethernet WOL (Wake On Lan) *magic packet* using the *ether-wake* application. It is important to point out that not all compute nodes will necessarily be in the same network, so the Power Management module must choose the correct network interface when sending the *magic packet*. This is configured in the dashboard. In order to shutdown a node, this can be done by simply executing the command *poweroff*. Another important remark is that for WOL to work, it must be enabled in the Ethernet

 ISBN: 978-84-616-9216-3

card, or it will ignore the packet. In order to assure that a powered off host can be powered on again, prior to each power off, the *ethtool* is used to enable WOL. If the host has an IPMI card the Power Management module can use it to power it on/off. This is done using tools such as *ipmiutil*.

2.3 Knowledge-based System

The key component of this architecture is a KBS implementing the decision-making mechanism that determines how many of the cluster resources must be on at every moment. At the core of the KBS, a set of if-the rules govern the behavior of the Power Management module. In this work, the same set of rules proposed in [10] are used, as mentioned before. These rules depend on a set of configuration parameters that are arguably flexible enough to match most of the desired cluster behaviours. This Knowledge-based System system can be expressed as:

- **if** $s_{running} + s_{starting} < s_{min}$ **then** power on $(s_{min} - (s_{running} + s_{starting}))$ slots
- **if** $t_{avg} > t_{max}$ or $n_{queued} > n_{max}$ **then** power on 1 slot
- **if** $t_{avg} < t_{min}$ or $n_{queued} < n_{min}$ **then** power off 1 slot
- **for each** h in *hosts* **do**
  if $i_h > i_{max}$ **then** power off host h

Where $s_{running}$ and $s_{starting}$ are the number of slots currently running and starting. s_{min} is the minimum number of slots required to run each of the queued jobs, that is, the maximum requested slots of an individual job among the queued ones. s_{total} are the cluster total slots (running and powered off). t_{avg} is the average waiting time for the queued jobs, and t_{max} and t_{min} are, respectively, the maximum and minimum average waiting time for the queued jobs. n_{queued} is the number of queued jobs, and n_{max} and n_{min} are the maximum and minimum number of queued jobs before an action is performed. Finally, i_h is the time that the host h has been at idle state and i_{max} is the maximum time that a host can be at idle state.

A particular instance of the Knowledge-based System can, therefore, be expressed as a combination of five parameters: $(t_{min}, t_{max}, n_{min}, n_{max}, i_{max})$.

2.4 Node selection

Once determined how many slots must be powered on/off, the next step is determine which specific nodes will be reconfigured. It is important to remark that only idle nodes would be powered off. The selection process involves two values: the node efficiency and the node timestamp of the last timed out.

 ISBN: 978-84-616-9216-3

The first one is calculated as $\frac{GFLOPS}{Watts}$, and the latter indicates the time of the last failure to power on/off upon request. In the first place, hosts are split by whether they succeeded or failed to comply with the last order. Those that succeeded are sorted according to their efficiency so that powered-on nodes are the most efficient and powered-off nodes are the least efficient ones. Conversely, those that failed are sorted according to the timestamps of their failures; those with the earliest values are always chosen. This mechanism allows the system to continuously iterate through the potentially malfunctioning nodes, thus increasing the possibility of finding a repaired one.

3 Evolutionary learning for multicriteria decision making

As mentioned before, the advantage of the Knowledge-based System detailed earlier is the ability to adapt to any desired working mode for the cluster due to the many configuration parameters that rule its operation. However, this ability to adapt comes with the problem of actually finding the right set of values to match the desired working mode. Firstly, the huge amount of value combinations makes an exhaustive search infeasible. Secondly, the optimal configuration involves, as many real world problems, multiple conflicting objectives instead of a single one. Because of this, there is not optimal solution but rather a set of optimal solutions (known as Pareto-optimal solutions or the Pareto Efficient Frontier) [9].

Multiobjective evolutionary algorithms (MOEAs) are widely regarded as an efficient method for finding Pareto-optimal solutions, from which an expert human can pick the preferred one [26]. In our research the chosen MOEA is the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [9], through its implementation in the MOEA Framework[4]. Every individual solution $(t_{min}, t_{max}, n_{min}, n_{max}, i_{max})$ that the algorithm finds, is evaluated by a running a simulation of the cluster with a determined workload so that the values for the multiple objectives are calculated. A description of NSGA-II can be found elsewhere and will not be repeated here; in the following of this section the definition of the specific multiobjective fitness function used in the problem at hand is given.

The fitness function consists in three conflicting criteria: the quality of service, the energy saved and the number of node reconfigurations. For a given set of n jobs, where the j-th job ($j = 1 \dots n$) is scheduled to start at time tsch_j, but effectively starts at time ton_j and stops at time toff_j, the quality of service in a HPC cluster reflects the amount of time that each job has to wait before is assigned its requested resources. Once the job starts its execution, it will not be halted, thus we focus only on its waiting time. Because jobs do not last the same amount of time, their waiting in the queue is better expressed as a ratio considering their execution time. Finally, due to the potential existence of outlier values, the 90 percentile is used instead of average:

[4]MOEA Framework, a Java library for multiobjective evolutionary algorithms, http://www.moeaframework.org/

 ISBN: 978-84-616-9216-3

$$\text{QoS} = \min\left\{p : ||\{j \in 1\ldots n : \frac{\text{ton}_j - \text{tsch}_j}{\text{toff}_j - \text{ton}_j} \leq p\}|| > 0.9\, n\right\} \quad (1)$$

where $||A||$ is the cardinality of the set A.

The energy saved is measured as the sum of the amount of seconds that each node has been powered off. Let c be the number of nodes, let state(i,t) be 1 if the i-th node ($i = 1\ldots c$) is powered at time t and 0 otherwise. Lastly, let the time scale be the lapse between tini=$\min_j\{\text{sch}_j\}$ and tend= $\max_j\{\text{toff}_j\}$. Then,

$$\text{Energy saved} = c \cdot (\text{tend} - \text{tini}) - \sum_{i=1}^{c} \int_{\text{tini}}^{\text{tend}} \text{state}(i,t)\text{d}t. \quad (2)$$

The node reconfigurations is the number of times that a node has been powered on or off. Let nd(i) the number of discontinuties of the function state(i,t) in the time interval $t \in (\text{tini}, \text{tend})$:

$$\text{Reconfigured nodes} = \sum_{i=1}^{c} \text{nd}(i) \quad (3)$$

4 Experimental results

The experimental setup is based on actual workloads from the Scientific Modelling Cluster of the University of Oviedo spanning 22 months, with a total of 2907 jobs. For both training and testing, a cluster simulator has been developed so that every model can be evaluated in the three criteria described in the previous section.

Three solutions have been tested using this simulator and the workloads: a) a basic model, b) the rule model proposed in [10], with its parameters manually configured by the administrator, and c) the learning mechanism proposed in this paper, using a NSGA-II algorithm. The holdout method was used for validation, with a 70-30% split in training and test.

The administrator preferences for the experiment are based upon a lexicographic ordering of the three criteria: the administrator always seeks the best QoS and the amount of energy saved is used only to break ties in QoS. In turn, the number of reconfigurations also serves to break ties in QoS and energy saving.

First, the basic model (labelled "Single rule" in Tables 1 and 2) consists on the allocation of as many compute node slots as are required to run all queued jobs, shutting down every idle node whenever the decision mechanism is triggered. Second, five different manual configurations were tested for the model in [10], intended to give different weights to QoS, energy and reconfigurations. Third, the machine learning approach (labelled "Rules NSGA-II") was applied to the same data. As shown in the aforementioned Tables 1 and 2, none of

the manual configurations neither the basic model was competitive with the machine learning approach. The experimentation shows that finding manually a suitable configuration for the multi-rule model is an infeasible task due to the large number of combinations.

Lastly, the Pareto Efficient Frontier obtained in this experiment is represented in Figure 2. The chosen configuration, marked with a black dot in the figure, achieves optimal QoS and also saves energy while keeping acceptable node reconfigurations, thus complying with the previously declared preferences. Observe that many different balances between energy consumption, QoS and reconfigured nodes can be obtained from this set of solutions, and also that none of the manually found sets of parameters is part of the set of Pareto-optimal configurations.

	Training set		
	QoS	Energy saved(s)	Reconfigurations
Single rule	157.80	1.26E+09	2755
Rules (0, 60, 0, 5, 3600)	112.00	1.29E+09	2047
Rules (0, 300, 0, 10, 3600)	184.10	1.29E+09	2023
Rules (0, 60, 0, 5, 7200)	103.30	1.29E+09	1945
Rules (0, 60, 0, 0, 14400)	93.66	1.28E+09	1845
Rules NSGA-II	0.00	8.54E+08	81

Table 1: Experiment results for the training set

	Test set		
	QoS	Energy saved(s)	Reconfigurations
Single rule	80.16	4.22E+08	2504
Rules (0, 60, 0, 5, 3600)	48.62	4.25E+08	1538
Rules (0, 300, 0, 10, 3600)	77.43	4.26E+08	1512
Rules (0, 60, 0, 5, 7200)	22.34	4.23E+08	1386
Rules (0, 60, 0, 0, 14400)	2.92	4.19E+08	1216
Rules NSGA-II	0.00	1.88E+08	47

Table 2: Experiment results for the test set

5 Concluding remarks and future work

An evolutionary learning algorithm has been designed that is able to optimize the parameters defining the rules in the KBS that drives the Power Management module of a HPC cluster. The new procedure has been tested with actual workloads captured at the Scientific Modelling Cluster at Oviedo University. It has been found that expert knowledge is not

 ISBN: 978-84-616-9216-3

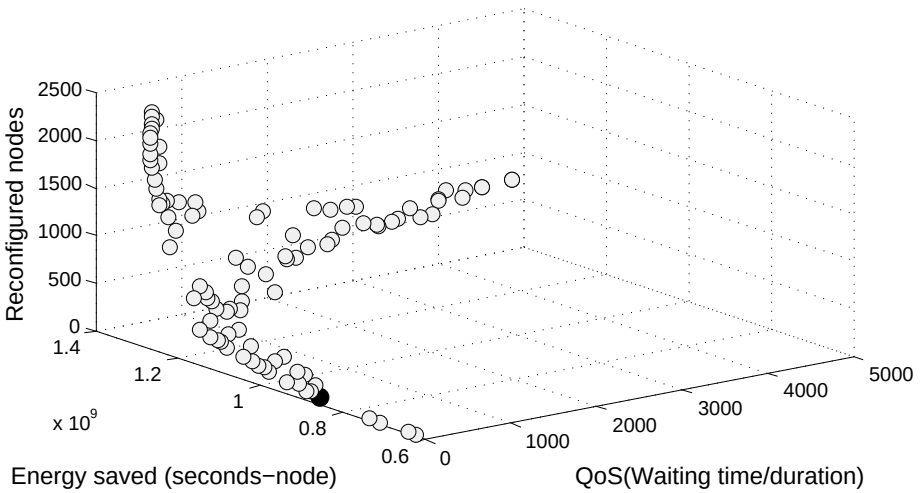


Figure 2: Pareto Efficient Frontier obtained in the experiment

enough for fine-tuning this system; the learning system was able to produce a combination of parameters that improved the initial solution in the three criteria, at the same time: QoS, energy saving and node reconfiguration.

One might wonder whether this approach is general enough for being applied to different scenarios, and what the expected gain would be in those cases. Further work is needed to provide a sound answer to this question. On the one hand, it is clear that the KBS is highly dependent on the expected profile of the workload. On the other hand, for those cases where the load does not follow a regular pattern, the improvement over the simpler schemes might not be relevant enough.

Lastly, it is remarked that the structure of the rule base is not currently part of the learning process. Also in future works, different parametric definitions of the rule base will be explored, including an extended learning algorithm that not only tunes the parameters defining the rules but fully learns the linguistic definition of the rule base.

Acknowledgements

This work has been partially supported by "Ministerio de Economía y Competitividad" from Spain/FEDER under grants TEC2012-38142-C04-04 and TIN2011-24302.

References

[1] P. Alonso, R. M. Badia, J. Labarta, M. Barreda, M. F. Dolz, R. Mayo, E. S. Quintana-Orti, and R. Reyes, *Tools for Power-Energy Modelling and Analysis of Parallel Scientific Applications*, in 2012 41st International Conference on Parallel Processing, IEEE, Sept. 2012, pp. 420–429.

[2] F. Alvarruiz, C. de Alfonso, M. Caballer, and V. Hern'ndez, *An Energy Manager for High Performance Computer Clusters*, in 2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications, IEEE, July 2012, pp. 231–238.

[3] C. Bash and G. Forman, *Cool job allocation: measuring the power savings of placing jobs at cooling-efficient locations in the data center*, USENIX Association, June 2007, p. 29.

[4] J. L. Berral, I. n. Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres, *Towards energy-aware scheduling in data centers using machine learning*, in Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking - e-Energy '10, New York, New York, USA, Apr. 2010, ACM Press, p. 215.

[5] R. Buyya, H. Jin, and T. Cortes, *Cluster computing*, Future Generation Computer Systems, 18 (2002), pp. v–viii.

[6] Y. Cheng and Y. Zeng, *Automatic Energy Status Controlling with Dynamic Voltage Scaling in Power-Aware High Performance Computing Cluster*, in 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies, IEEE, Oct. 2011, pp. 412–416.

[7] G. L. T. Chetsa, L. Lefrvre, J.-M. Pierson, P. Stolf, and G. Da Costa, *A Runtime Framework for Energy Efficient HPC Systems without a Priori Knowledge of Applications*, in 2012 IEEE 18th International Conference on Parallel and Distributed Systems, IEEE, Dec. 2012, pp. 660–667.

[8] R. Das, J. O. Kephart, C. Lefurgy, G. Tesauro, D. W. Levine, and H. Chan, *Autonomic multi-agent management of power and performance in data centers*, (2008), pp. 107–114.

[9] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, *A fast and elitist multiobjective genetic algorithm: NSGA-II*, IEEE Transactions on Evolutionary Computation, 6 (2002), pp. 182–197.

 ISBN: 978-84-616-9216-3

[10] M. F. Dolz, J. C. Fernández, S. Iserte, R. Mayo, E. S. Quintana-Ortí, M. E. Cotallo, and G. Díaz, *EnergySaving Cluster experience in CETA-CIEMAT*, in 5th Iberian GRID Infrastructure conference, Santander, 2011.

[11] M. Ebbers, Mike Archibald, C. F. F. da Fonseca, M. Griffel, V. Para, and M. Searcy, *Smarter Data Centers: Achieving Greater Efficiency*, tech. report, IBM Redpaper, 2011.

[12] EIA, *Electric Power Monthly - Energy Information Administration.*

[13] E. N. Elnozahy, M. Kistler, and R. Rajamony, *Energy-efficient server clusters*, (2002), pp. 179–197.

[14] Emerson Network Power, *Energy Logic: Reducing Data Center Energy Consumption by Creating Savings that Cascade Across Systems*, tech. report, 2009.

[15] Eurostat, *Electricity and natural gas price statistics - Statistics Explained*, 2013.

[16] V. W. Freeh and D. K. Lowenthal, *Using multiple energy gears in MPI programs on a power-scalable cluster*, in Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming - PPoPP '05, New York, New York, USA, June 2005, ACM Press, p. 164.

[17] V. W. Freeh, D. K. Lowenthal, F. Pan, N. Kappiah, R. Springer, B. L. Rountree, and M. E. Femal, *Analyzing the Energy-Time Trade-Off in High-Performance Computing Applications*, IEEE Transactions on Parallel and Distributed Systems, 18 (2007), pp. 835–848.

[18] D. F. Garcia, J. Entrialgo, J. Garcia, and M. Garcia, *A self-managing strategy for balancing response time and power consumption in heterogeneous server clusters*, in 2010 International Conference on Electronics and Information Engineering, vol. 1, IEEE, Aug. 2010, pp. V1–537–V1–541.

[19] Gartner, *Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO2 Emissions*, 2007.

[20] R. Ge, X. Feng, W.-c. Feng, and K. W. Cameron, *CPU MISER: A Performance-Directed, Run-Time System for Power-Aware Clusters*, in 2007 International Conference on Parallel Processing (ICPP 2007), IEEE, Sept. 2007, pp. 18–18.

[21] R. Haring, *The Blue Gene/Q Compute Chip*, tech. report, IBM Corporation, 2011.

[22] C.-H. Hsu and W.-c. Feng, *A Power-Aware Run-Time System for High-Performance Computing*, in ACM/IEEE SC 2005 Conference (SC'05), IEEE, 2005, pp. 1–1.

[23] C.-H. Hsu and U. Kremer, *The design, implementation, and evaluation of a compiler algorithm for CPU energy reduction*, ACM SIGPLAN Notices, 38 (2003), p. 38.

[24] S. Huang and W. Feng, *Energy-Efficient Cluster Computing via Accurate Workload Characterization*, in 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, IEEE, 2009, pp. 68–75.

[25] IBM Systems and Technology Group, *IBM System Blue Gene/Q - DCD12345USEN.pdf*, tech. report, IBM, Somers, NY, 2011.

[26] M. Jensen, *Reducing the Run-Time Complexity of Multiobjective EAs: The NSGA-II and Other Algorithms*, IEEE Transactions on Evolutionary Computation, 7 (2003), pp. 503–515.

[27] W. Lang, J. M. Patel, and J. F. Naughton, *On energy management, load balancing and replication*, ACM SIGMOD Record, 38 (2010), p. 35.

[28] D. Li, D. S. Nikolopoulos, K. Cameron, B. R. de Supinski, and M. Schulz, *Power-aware MPI task aggregation prediction for high-end computing systems*, in 2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS), IEEE, 2010, pp. 1–12.

[29] M. Lim, V. Freeh, and D. Lowenthal, *Adaptive, Transparent Frequency and Voltage Scaling of Communication Phases in MPI Programs*, in ACM/IEEE SC 2006 Conference (SC'06), IEEE, Nov. 2006, pp. 14–14.

[30] R. M. Llamas, D. F. Garcia, and J. Entrialgo, *A Technique for Self-Optimizing Scalable and Dependable Server Clusters under QoS Constraints*, in 2012 IEEE 11th International Symposium on Network Computing and Applications, IEEE, Aug. 2012, pp. 61–66.

[31] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, *Load balancing and unbalancing for power and performance in cluster-based systems*, in Workshop on compilers and operating systems for low power, vol. 180, Barcelona, Spain, 2001, pp. 182–195.

[32] S. Schubert, D. Kostic, W. Zwaenepoel, and K. G. Shin, *Profiling Software for Energy Consumption*, in 2012 IEEE International Conference on Green Computing and Communications, IEEE, Nov. 2012, pp. 515–522.

[33] G. Tang, Q. and Gupta, S. K S and Varsamopoulos, *Energy-Efficient Thermal-Aware Task Scheduling for Homogeneous High-Performance Computing Data Centers: A Cyber-Physical Approach*, IEEE Transactions on Parallel and Distributed Systems, 19 (2008), pp. 1458–1472.

 ISBN: 978-84-616-9216-3

[34] The Economist Intelligence Unit, *IT and the environment A new item on the CIOs agenda?*, tech. report, The Economist, 2007.

[35] U.S. Environmental Protection Agency, *Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431*, tech. report, ENERGY STAR Program, 2007.

[36] G. L. Valentini, W. Lassonde, S. U. Khan, N. Min-Allah, S. A. Madani, J. Li, L. Zhang, L. Wang, N. Ghani, J. Kolodziej, H. Li, A. Y. Zomaya, C.-Z. Xu, P. Balaji, A. Vishnu, F. Pinel, J. E. Pecero, D. Kliazovich, and P. Bouvry, *An overview of energy efficiency techniques in cluster computing systems*, Cluster Computing, 16 (2011), pp. 3–15.

[37] C. Xian, Y.-H. Lu, and Z. Li, *A programming environment with runtime energy characterization for energy-aware applications*, in Proceedings of the 2007 international symposium on Low power electronics and design - ISLPED '07, New York, New York, USA, Aug. 2007, ACM Press, pp. 141–146.

[38] Z. Xue, X. Dong, S. Ma, S. Fan, and Y. Mei, *An Energy-Efficient Management Mechanism for Large-Scale Server Clusters*, in The 2nd IEEE Asia-Pacific Service Computing Conference (APSCC 2007), IEEE, Dec. 2007, pp. 509–516.

[39] F. Yeo, CheeShin and Buyya, Rajkumar and Pourreza, Hossein and Eskicioglu, Rasit and Graham, Peter and Sommers, *Cluster Computing: High-Performance, High-Availability, and High-Throughput Processing on a Network of Computers*, in Handbook of Nature-Inspired and Innovative Computing, A. Zomaya, ed., Springer US, 2006, pp. 521–551.

[40] Z. Zong, M. Nijim, A. Manzanares, and X. Qin, *Energy efficient scheduling for parallel applications onmobileclusters*, Cluster Computing, 11 (2007), pp. 91–113.

[41] Z. Zong, X. Ruan, A. Manzanares, K. Bellam, and X. Qin, *Improving Energy-Efficiency of Computational Grids via Scheduling*, in Handbook of Research on P2P and Grid Systems for Service-Oriented Computing, N. Antonopoulos, G. Exarchakos, M. Li, and A. Liotta, eds., IGI Global, Jan. 2010, ch. 22.

 ISBN: 978-84-616-9216-3