Proceedings of the 2013
International Conference on
Computational and Mathematical
Methods in Science and Engineering

Almería, Spain

June 24-27, 2013

CMMSE

VOLUME I

Editors: Ian Hamilton & Jesús Vigo-Aguiar

Associate Editors:

H. Adeli, P. Alonso, M.T. De Bustos, M. Demiralp, J.A. Ferreira, A. Q. M. Khaliq,

J.A. López-Ramos, P. Oliveira, J.C. Reboredo, M. Van Daele,

E. Venturino, J. Whiteman, B. Wade

Proceedings of the 2013 International Conference on Computational and Mathematical Methods in Science and Engineering

Cabo de Gata, Almería, Spain

June 24-27, 2013

CMMSE
Computational and Mathematical
Methods in Science and Engineering

Editors

I. P. Hamilton & J. Vigo-Aguiar

Associate Editors

H. Adeli, P. Alonso, M.T. De Bustos,
M. Demiralp, J.A. Ferreira, A. Q. M. Khaliq,
J.A. López-Ramos, P. Oliveira, J.C. Reboredo,
M. Van Daele, E. Venturino, J. Whiteman, B. Wade

ISBN 978-84-616-2723-3

Printed on acid-free paper

Preface

It is our great pleasure to present the proceedings of the **13th International Conference on Computational and Mathematical Methods in Science and Engineering** (CMMSE 2013), at Cabo de Gata, Almería (Spain), 24-27 June 2013, comprised of the extended abstracts of the conference presentations.

Since its inception in Milwaukee in 2000, CMMSE has provided a stimulating annual forum for researchers, from a wide range of disciplines, who have found in CMMSE a fruitful arena in which to disseminate their contributions to the research community. Researchers also benefit from being at the crossroads of computational and mathematical methods in a wide variety of research areas. Encouraging the development of the new computational and mathematical methods increasingly demanded by diverse disciplines continues to be the cornerstone of the conference.

Interest in understanding the behaviour of complex systems and phenomena is growing since this essential for the technological development of our society. The resolution of many problems at the theoretical and practical level in science, engineering, economics, and finance requires the intensive development of computational and mathematical methods, which have thus become essential research tools. CMMSE 2013 covers all the computational and mathematical fields, providing specific responses for specific fields and describing up-to-date developments to an expert audience. In addition, mini-symposia and special sessions cover a wide range of specialized topics.

New large-scale problems that arise in fields like bioinformatics, computational chemistry, and astrophysics are considered in a high-performance computing session, whereas mathematical problems related to internet security are considered in another session. Likewise, analytical, numerical and computational aspects of partial differential equations in life and materials science are considered in a specific mini-symposium. The computational finance session covers 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. The symposium on new educational methodologies supported by new technologies offers a forum for discussion of the growing impact of new technologies on teaching, and the development of new tools to increase learning efficiency. Flow-modelling of particles with motivated behaviour in complex networks, applied to traffic flows, pedestrian flows, ecology, etc., are presented in the symposium on mathematical models and information-intelligent transport systems. Recent techniques to solve various types of optimization problems in engineering are presented in the session on computational methods for linear and nonlinear optimization, while numerical methods for solving nonlinear problems are presented in another specific session. Recent theoretical and applied mathematical developments related to cryptography and codes, bio-mathematics, combinatorial optimization algorithms, and algebraic analysis are presented in other specific sessions. Novel methods for the approximation of univariate and multivariate functions, the solution of ordinary and partial differential equations, and the decomposition of multivariate arrays are presented in two mini-symposia. Another mini-symposium covers new computational methods for improving computed tomography reconstruction quality and speed. Finally, special sessions cover topics related to industrial mathematics, *ab initio* materials and simulation, 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, engineering, and computational finance. 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 2013. Finally, we would like to thank the participants because, without their interest and enthusiasm, the conference would not have been possible.

These proceedings, comprised of the extended abstracts of the conference presentations, 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.

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

Cabo de Gata, Almería (Spain), 20 June 2013

I. P. Hamilton, J. Vigo-Aguiar, H. Adeli, P. Alonso,
M.T. De Bustos, M. Demiralp, J.A. Ferreira, A. Q. M. Khaliq,
J.A. López-Ramos, P. Oliveira, J.C. Reboredo,
M. Van Daele, E. Venturino, J. Whiteman, B. Wade

CMMSE 2013 Mini-symposia

Session Title	Organizers
High Performance Computing (HPC)	Enrique S. Quintana Ortí & G. Ester Martín Garzón
P.D.E.'S in Life and Material Sciences	Paula Oliveira & J.A. Ferreira
Computational Finance	Abdul Khaliq & Juan C. Reboredo
New Educational Methodologies Supported by New Technologies	José A. Piedra
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
Crypto & Codes	Juan Antonio López-Ramos & İrfan Şiap
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

CMMSE 2013 Special Sessions

Session Title	Organizers
Industrial Mathematics	Bruce A. Wade
Ab initio materials & simulation	Ian Hamilton
Computational Discrete Mathematics	José Carlos Valverde Fajardo
Numerical solution of differential equations	J. Vigo Aguiar & Marnix Van Daele
Computational Methods in Tomography	J. Román Bilbao Castro, G. Ester Martín Garzón & J. J. Fernández Rodriguez

Acknowledgements

We would like to express our gratitude to Universidad de Almeria, ASAC communications & Patronato de la Alhambra y Generalife our sponsors, for its assistance.

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

- Pedro Alonso - *Universidad de Oviedo - Gijón, Spain*
- José A. Álvarez-Bermejo – *Universidad de Almería, Spain*
- M. Teresa de Bustos - *Universidad de Salamanca, Spain*
- Antonio Fernández - *Universidad de Salamanca, Spain*
- Pedro Gonzalez Rodelas - *Universidad de Granada, Spain*
- Juan Antonio López Ramos - *Universidad de Almería, Spain*
- Raquel Cortína Pajarón - *Universidad de Oviedo -Gijón, Spain*
- Antonio Sánchez Martín - *Universidad de Salamanca, Spain*

CMMSE 2013 Plenary Speakers

- Abdul Q. M. Khaliq - Middle Tennessee State University, **USA**
- Ezio Venturino - University of Torino, **Italy**
- José M. Soler - Univ. Autónoma de Madrid , **Spain**
- John Whiteman, - Brunel University, London, **UK**
- Marnix Van Daele - University of Gent, **Belgium**
- Metin Demiralp - Turkish Academy of Sciences, **Turkish**
- Mikel Luján - Manchester University, **UK**
- Nico Stollenwerk - University Lisbon, **Portugal**

Volume I

Contents:

Volume I

Contents:

Volume II

Contents:

Volume III

Contents:

Volume IV

Contents:

Volume V

.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Determinants and inverses of nonsingular pentadiagonal 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

For nonsingular $n \times n$ ($n \geq 6$) pentadiagonal matrices $\mathbf{P}$ having nonzero entries on its second subdiagonal, we propose a procedure for computing both the determinant $\det \mathbf{P}$ in $O(n)$ times, and accurate information for obtaining the inverse $\mathbf{P}^{-1}$ in $O(n^2)$ times. In the general nonsingular case, $n \geq 5$, a suitable decomposition of $\mathbf{P}$, as a product of two nonsingular upper Hessenberg matrices, allows us another procedure for obtaining both $\det \mathbf{P}$ and $\mathbf{P}^{-1}$ taking advantage of such low rank structured matrices.

Key words: Determinant, inverse, matrix computations, pentadiagonal matrix.

1 Introduction

Nonsingular pentadiagonal matrices of a finite order n, $\mathbf{P} = \{p_{i,j}\}_{1 \leq i,j \leq n}$ (with $p_{i,j} = 0$ for $|i-j| > 2$) have a role in current methods of the numerical analysis. They frequently arise in ODEs, PDEs, interpolation and spline problems [4], boundary value problems, BVP, involving fourth order derivatives. Also, pentadiagonal matrices appear and in finer approximations of second order derivatives. Gauss-Jordan methods with partial pivoting are usually handled in the inversion of such matrices. However, these methods can destroy the special structure and sparsity of the pentadiagonal matrices. Hence, computational techniques based on the low rank structure of the pentadiagonal matrices are of interest.

Particular algorithms for the inversion of $\mathbf{P}$ are known. In the sequential line, a procedure for the inverse of $\mathbf{P}$ was provided in [3], with the condition in the entries $p_{i,j} \neq 0$ for $i - j = 2$, or $j - i = 2$, in $O(n^2)$ times. Another procedure with complexity $O(n^2)$ was proposed for pentadiagonal matrices having an LU (Doolittle) factorization, [7].

Fast numerical algorithms for the determinant of pentadiagonal matrices $\mathbf{P}$ are required to test efficiently the existence of unique solutions of the PDEs, and also for inverse construction methods of symmetric pentadiagonal Toeplitz matrices. Some results with complexity $O(n)$ have been obtained for the determinant of nonsingular pentadiagonal matrices $\mathbf{P}$, [2, 5, 6]. As a continuation of this line, we propose a fast and accurate computation, also with complexity $O(n)$, for the determinant of a pentadiagonal matrix having nonzero entries in its second subdiagonal, currently used in numerical analysis. In addition, all the accurate information about the inverse $\mathbf{P}^{-1}$ is obtained with complexity $O(n^2)$.

The computation of the determinant and the inverse of any nonsingular pentadiagonal matrix $\mathbf{P}$ taking advantage of its special low rank structure, and without conditions on its entries, is an open question. We propose here a simple factorization for the general nonsingular case, where the pentadiagonal matrix $\mathbf{P}$ can be decomposed as a product of two suitable upper Hessenberg matrices. It provides us the determinant and the inverse of $\mathbf{P}$ by exploiting the low rank structure of the Hessenberg matrices [1].

2 Pentadiagonal matrices having nonzero entries in its second subdiagonal.

The 2×2 block structure $\mathbf{P} = \left(\begin{array}{c|c} \mathbf{P}_{11} & \mathbf{0}_2 \\ \hline \mathbf{U} & \mathbf{P}_{22} \end{array}\right)$, for a $n \times n$ ($n \geq 6$) nonsingular pentadiagonal matrix is assumed. $\mathbf{P}_{11}$ and $\mathbf{P}_{22}$ are matrices of order $2 \times n-2$ and $n-2 \times 2$, respectively. Matrix $\mathbf{0}_2$ is the zero matrix of order 2. The $n-2 \times n-2$ matrix $\mathbf{U}$ is nonsingular upper triangular. The transposed partitioning for its inverse is known, $\mathbf{P}^{-1} = \left(\begin{array}{c|c} -\mathbf{U}^{-1}\mathbf{P}_{22}\mathbf{M}_{21} & \mathbf{U}^{-1} + \mathbf{U}^{-1}\mathbf{P}_{22}\mathbf{M}_{21}\mathbf{P}_{11}\mathbf{U}^{-1} \\ \hline \mathbf{M}_{21} & -\mathbf{M}_{21}\mathbf{P}_{11}\mathbf{U}^{-1} \end{array}\right)$, with $\mathbf{M}_{21} = \frac{1}{\det \mathbf{P}} \begin{pmatrix} C_{1,n-1} & C_{2,n-1} \\ C_{1,n} & C_{2,n} \end{pmatrix}$, and the given $C_{i,j}$ are cofactors of $\mathbf{P}$. Therefore, $\mathbf{P}^{-1}$ can be seen as:

$$\mathbf{P}^{-1} = \begin{pmatrix} -\mathbf{U}^{-1}\mathbf{P}_{22} \\ \mathbf{I}_2 \end{pmatrix} \mathbf{M}_{21} \begin{pmatrix} \mathbf{I}_2 & -\mathbf{P}_{11}\mathbf{U}^{-1} \end{pmatrix} + \left(\begin{array}{c|c} \mathbf{0}_{n,2} & \mathbf{U}^{-1} \\ \hline \mathbf{0}_2 & \mathbf{0}_{2,n} \end{array}\right), \tag{1}$$

a rank two perturbation of a strictly upper triangular matrix. Just consider as, given $\mathbf{P}$, the information required for the inversion of this class of pentadiagonal matrices is contained in the matrices $\mathbf{M}_{21}$ and $\mathbf{U}^{-1}$. From these matrices, we can obtain the inverse using (1).

2.1 Computing the determinant in $O(n)$ times

Proposition 1 *Let* $\mathbf{P}$ *a* $n \times n$ ($n \geq 6$) *nonsingular pentadiagonal matrix having nonzero entries on its second subdiagonal. Then,* $\det \mathbf{P}$ *can be computed in* $O(n)$ *times by*

$$\det \mathbf{P} = \left(\prod_{k=1}^{n-2} p_{k+2,k}\right) \det \begin{pmatrix} C^*_{1,n-1} & C^*_{2,n-1} \\ C^*_{1n} & C^*_{2n} \end{pmatrix}, \tag{2}$$

where the C^*_{ji} *are cofactors of the matrix* $\mathbf{P}^* = \mathbf{P} \cdot \mathbf{diag}\left(\frac{1}{p_{31}}, \frac{1}{p_{42}}, \cdots, \frac{1}{p_{n,n-2}}, 1, 1\right)$.

Proof. It is a consequence of the formula $\det \mathbf{P} = \left(\prod_{k=1}^{n-2} p_{k+2,k}\right) \cdot X_n^{'}$ given in [3]. Here, we have chosen the transposed matrix because the determinant is the same. From the recurrences (5) and (6) given in [3] taking the transposes, we observe that $X_n^{'} = \det \begin{pmatrix} (-1)^n C^*_{1,n} & (-1)^{n-1} C^*_{1,n-1} \\ (-1)^n C^*_{2n} & (-1)^{n-1} C^*_{2,n-1} \end{pmatrix} = \det \begin{pmatrix} C^*_{1,n-1} & C^*_{2,n-1} \\ C^*_{1n} & C^*_{2n} \end{pmatrix}$, where the given cofactors of $\mathbf{P}^*$ are involved.

Note also that $\det \mathbf{P}^* = \det \begin{pmatrix} C^*_{1,n-1} & C^*_{2,n-1} \\ C^*_{1n} & C^*_{2n} \end{pmatrix}$, where the cofactors are related with determinants of sparse upper Hessenberg matrices and with ones on their subdiagonal. Therefore, such determinants can be computed with complexity $O(n)$. ■

Given $\mathbf{P}$ we require as main cost $13n + O(1)$ quotients and products for obtaining $\det \mathbf{P}$; $5n$ operations for computing $\mathbf{P}^*$ and $8n$ operations for computing the cofactors from (2).

In Table 1 we compare with the built-in routine of *Matlab*® commercial package *det*(), and the Sogabe algorithm [5], for a current matrix $\mathbf{P}$ with entries $p_{i,j} = 1$, for $|i - j| \leq 2$, and $p_{i,j} = 0$ otherwise. Sogabe algorithm breaks down because some principal submatrices are singular. Our procedure also works with singular pentadiagonal matrices with non zero entries in the second subdiagonal.

2.2 Computing the inverse in $O(n^2)$ times

The procedure derived from Proposition 1 can be used to obtain the matrices $\mathbf{M}_{21}$ and $\mathbf{U}^{-1}$, containing all the information of the inverse matrix $\mathbf{P}^{-1}$. It is no difficult to observe that $\mathbf{M}_{21} = \frac{1}{\det \mathbf{P}} \begin{pmatrix} C_{1,n-1} & C_{2,n-1} \\ C_{1n} & C_{2n} \end{pmatrix} = \frac{1}{\det \mathbf{P}^*} \begin{pmatrix} C^*_{1,n-1} & C^*_{2,n-1} \\ C^*_{1n} & C^*_{2n} \end{pmatrix}$. Thus we can obtain $\mathbf{M}_{21}$ with complexity $O(n)$. Also $\mathbf{U}^{-1}$ can easily be obtained from the matrix $\mathbf{P}^*(3:n, 1:n-2)$. By reason of the sparsity and the ones on the main diagonal of such matrix, we can obtain $\mathbf{U}^{-1}$ with reasonable accuracy in $O(n^2)$ times. That is, given $\mathbf{P}^*(3:n, 1:n-2)$ in $O(n)$ times, the main cost for obtaining $\mathbf{U}^{-1}$ is $2n^2 - 10n + 2$ products. With $\mathbf{M}_{21}$, $\mathbf{U}^{-1}$, and the matrix operations from (1), we supply the inverse in $O(n^2)$ times.

3 Nonsingular pentadiagonal matrices: the general setting.

For general $n \times n$ ($n \geq 5$) nonsingular pentadiagonal matrices, we assume a 2×2 block structure $\mathbf{P} = \left(\begin{array}{c|c} \mathbf{R} & 0 \\ \hline \mathbf{H} & \mathbf{C} \end{array}\right)$. Here $\mathbf{R}$ is a $1 \times n-1$ row, and $\mathbf{C}$ a $n-1 \times 1$ column matrix. $\mathbf{H}$ is an $n-1 \times n-1$ reduced upper Hessenberg matrix. The unreduced case is also applicable, but it is efficiently handled in Section 2.

Order	Matlab® ET	Proposed ET	Matlab® value	Proposed value	Sogabe
27	0.96e-04	0.68e-04	0	0	NaN
34	2.33e-04	1.22e-04	0	0	NaN
41	1.75e-04	0.90e-04	1	1	NaN
48	2.57e-04	1.22e-04	0	0	NaN
55	2.03e-04	0.89e-04	1	1	NaN

Table 1: Values given by the algorithms for elapsed times (ET) and the determinant of a pentadiagonal matrix $\mathbf{P}$ with entries $p_{i,j} = 1$, for $|i - j| \leq 2$, and $p_{i,j} = 0$ otherwise.

With $\mathbf{H}$ reduced, we take the factorizations $\mathbf{P} = \left(\begin{array}{c|c} 1 & \mathbf{0}_{n-1}^T \\ \hline \mathbf{0}_{n-1} & \mathbf{H} \end{array}\right)\left(\begin{array}{c|c} \mathbf{R} & 0 \\ \hline \mathbf{I}_{n-1} & \mathbf{H}^{-1}\mathbf{C} \end{array}\right)$ for $\mathbf{H}$ nonsingular, and $\mathbf{P} = \left(\begin{array}{c|c} 1 & \mathbf{0}_{n-1}^T \\ \hline \mathbf{0}_{n-1} & \mathbf{H}^* \end{array}\right)\left(\begin{array}{c|c} \mathbf{R} & 0 \\ \hline \mathbf{U}^* & \mathbf{H}^{*-1}\mathbf{C} \end{array}\right)$ for $\mathbf{H}$ singular. We can choose $\mathbf{H} = \mathbf{H}^* \cdot \mathbf{U}^*$ in the singular case, with $\mathbf{H}^*$ nonsingular with the same lower half and diagonal than $\mathbf{H}$. Then, the upper triangular matrix $\mathbf{U}^*$ must be singular. In both factorizations the component matrices are nonsingular upper Hessenberg matrices. From such decompositions based on upper Hessenberg matrices, the determinant and inverse of $\mathbf{P}$ can be computed taking advantage of the low rank structure of such matrices; see [1].

References

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

[2] Z. Cinkir, *An elementary algorithm for computing the determinant of pentadiagonal Toeplitz matrices*, J. Comp. Appl. Math. **236** (2012) 2298–2305.

[3] M. Elouafi, A.D. Aiat Hadj, *A fast numerical algorithm for the inverse of a tridiagonal and pentadiagonal matrix*, Appl. Math. Comp. **202** (2008) 441–445.

[4] J. M. McNally, *A fast algorithm for solving diagonally dominant symmetric pentadiagonal Toeplitz systems*, J. Comp. Appl. Math. **234** (2010) 995–1005.

[5] T. Sogabe, *A note on "A fast numerical algorithm for the determinant of a pentadiagonal matrix"*, Appl. Math. Comp. **201** (2008) 561–564.

[6] R. A. Sweet, *A Recursive Relation for the Determinant of a Pentadiagonal Matrix*, Comm. ACM **12** (1969) 330–332.

[7] X. L. Zhao, T. Z. Huang, *On the inverse of a general pentadiagonal matrix*, Appl. Math. Comp. **202** (2008) 639–646.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

A new tool for generating orthogonal polynomial sequences

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, 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

Adequate conditions, using a known result on a class of finite Hessenberg matrices, are here proposed to make available finite, and infinite, matrices as a rank one perturbation of strictly upper triangular matrices $UV + T$. Some characterizations on such matrices for generating orthogonal polynomial sequences are also considered.

Key words: Hessenberg matrix, inverse matrix, orthogonal polynomials, hermitian moment problem.

1 Introduction

The orthogonal polynomials, [7] are currently applied in many branches of science and engineering. They have a determinantal representation, involving particular Hessenberg matrices. A characterization for the nonsingular unreduced Hessenberg matrices in the finite case is related with the particular structure of its inverse matrix, [5, 6]. Such inverse is 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. Some conditions on their entries must be accomplished. In this work we want to obtain the unreduced Hessenberg matrix D, or the Jacobi matrix in the symmetric case, and related with sequence of orthogonal polynomials, by inverting an adequate matrix, without invoking the matrix derived from the dot product. The moment problem is no considered. Hence, we only need two adequate sequences $\{U\}$ and $\{V\}$, and an associated upper triangular matrix T, so that the matrix to be inverted is $UV + T$. From the characteristics of U, V, and T, not only algebraic conditions can be obtained, but some other analytical and topological properties. Therefore, we propose the new tool with some algebraic results, and some other possible connections remaining as open lines. Nevertheless, the final task could be about the conditions on $\{U\}$, $\{V\}$, and T, so that the inverse of the matrix $UV + T$ be a subnormal operator.

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]. We also recall that an upper Hessenberg matrix $H = (h_{ij})_{i,j=1}^{n}$ having nonzero entries in its subdiagonal, $h_{i+1,i} \neq 0$, and $i = 1, 2, ..., n-1$, is an unreduced upper Hessenberg matrix.

Lemma 1 *A nonsingular matrix $H = (h_{ij})_{1 \leq i,j \leq n}$ 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 strictly upper triangular having null entries in the main diagonal and nonzero entries in the superdiagonal, $t_{i,i+1} = \frac{1}{h_{i+1,i}} \neq 0$, $1 \leq i \leq n-1$.*

Proof. First, we assume that H is an unreduced upper Hessenberg matrix. A direct computation of its inverse H^{-1} using the cofactor matrix gives, for $i \geq j$,

$$\begin{aligned} b_{ij} &= \frac{\mathrm{Adj}(j,i)}{|H|} = \frac{(-1)^{i+j}}{|H|} \begin{vmatrix} H_{j-1} & D & E \\ 0 & F & G \\ 0 & 0 & H_{n-i}^{(i)} \end{vmatrix} = \frac{(-1)^{i+j}}{|H|} |H_{j-1}| [h_{j+1,j} \cdots h_{i,i-1}] |H_{n-i}^{(i)}| = \\ &= \frac{(-1)^{i-1}}{|H|} |H_{n-i}^{(i)}| \frac{1}{[h_{i+1,1} \cdots h_{n,n-1}]} \cdot (-1)^{j-1} |H_{j-1}| [h_{j+1,j} \cdots h_{n,n-1}], \end{aligned}$$

with $|H_{j-1}|$ the $j-1$ left principal minor and $|H_{n-i}^{(i)}|$ the $n-i$ right principal minor from the matrix H. Matrix F is a triangular one, containing on its diagonal entries from the subdiagonal of H. If we define

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

we have $b_{ij} = u_i v_j$, $i \geq j$. Taking the conventions $|H_0| = 1$ and $|H_0^{(n)}| = 1$, we observe that v_1 and u_n are nonzero entries. This fact about the lower half of B gives us the adequate structure of the inverse,

$$B = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix} + \begin{pmatrix} 0 & t_{12} & t_{13} & \cdots & t_{1n} \\ 0 & 0 & t_{23} & \cdots & t_{2n} \\ 0 & 0 & 0 & \cdots & t_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} = UV + T.$$

Hence,

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

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

The determinant of B gives, $|B| = v_1u_n \begin{vmatrix} u_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ u_2 & u_2v_2 & b_{23} & \cdots & b_{2n} \\ u_3 & u_3v_2 & u_3v_3 & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & v_2 & v_3 & \cdots & v_n \end{vmatrix}$.

Subtracting, in each row r_i of the previous matrix, the last row r_n by u_i, there results

$$|B| = v_1u_n \begin{vmatrix} 0 & b_{12} - u_1v_2 & b_{13} - u_1v_3 & \cdots & b_{1n} - u_1v_n \\ 0 & 0 & b_{23} - u_2v_3 & \cdots & b_{2n} - u_2v_n \\ 0 & 0 & 0 & \cdots & b_{3n} - u_3v_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & v_2 & v_3 & \cdots & v_n \end{vmatrix}.$$

Therefore,

$$|B| = (-1)^{n+1}v_1u_n \prod_{i=1}^{n-1}(b_{i,i+1} - u_iv_{i+1}) = v_1u_n \prod_{i=1}^{n-1}(-t_{i,i+1}) = \frac{v_1u_n}{\prod_{i=1}^{n-1}(-h_{i+1,i})}. \tag{2}$$

The last equality is obtained from the expression $b_{i,i+1} = u_iv_{i+1} + \frac{1}{h_{i+1,i}}$; see [2], Proposition 2. Hence, $t_{i,i+1} = \frac{1}{h_{i+1,i}} \neq 0$.

Now we assume that H is the inverse matrix of a nonsingular matrix $B = (b_{ij})$ as given in (1), with entries $b_{ij} = u_iv_j$, $i \geq j$, and $b_{ij} = u_iv_j + t_{i,j}$, $i < j$, with $1 \leq i, j \leq n$. Also, $t_{i,i+1} = \frac{1}{h_{i+1,i}} \neq 0$, $u_n \neq 0$, and $v_1 \neq 0$. We note from (2) that $|B|$ is independent of the entries b_{ij}, and $j - i \geq 2$. The adjoints of these entries are null values, and the nonsingular matrix $H = B^{-1}$ is upper Hessenberg. In addition, as B is nonsingular, its determinant $|B| \neq 0$, the $t_{i,i+1} = \frac{1}{h_{i+1,i}} \neq 0$, and H is unreduced. ■

An equivalent lemma can be obtained for the lower Hessenberg matrices.

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 known [5].

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

$$b_{ij} = u_i v_j \qquad \text{for } i \geq j, \qquad b_{ij} = w_i t_j \qquad \text{for } i \leq j,$$

and the entries u_1, v_n, w_n, and t_1 are nonzero.

Proof. As a tridiagonal matrix is at the same time lower and upper Hessenberg, this result is an immediate consequence from Lemma 1. ■

Trivially, $u_k v_k = w_k t_k$. If in addition the matrix is symmetric, $u_i = t_i$, and $v_j = w_j$.

Example 1 *From Lemma 1 on finite matrices, for the real symmetric tridiagonal matrix of an order n:*

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

with $a > 0$, we obtain the matrix B_n:

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

Because for n determined we expand the determinant $|J_n|$ by the first column, the involved determinants can easily be computed by the three-term recurrence:

$$|J_n| = b|J_{n-1}| - a^2|J_{n-2}|. \tag{3}$$

For the conditions of main interest, $b^2 - 4a^2 > 0$, we have

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

that can trivially be obtained using induction.

Some methods for inverting finite tridiagonal matrices are available; see e.g. [1, 3].

2 Orthogonal polynomials and Hessenberg matrices

Let μ be a finite and positive Borel measure on a bounded domain of the complex plane. If we define the moments as $c_{ij} = \int_{\text{supp}\mu} z^i \overline{z}^j d\mu(z)$, we have an hermitian definite positive (HDP) matrix $M = (c_{ij})_{i,j=0}^{\infty}$. If the measure lies on a subset of the real numbers, the moments are $S_{ij} = \int_{\text{supp}\mu} x^{i+j} d\mu(x)$, and $M = (S_{ij})_{i,j=0}^{\infty}$ is a Hankel matrix.

From a HDP matrix M, no necessarily a moment matrix, we can obtain a Hessenberg matrix D associated to M. The finite sections of D are $D_n = T_n^{-1} M_n' T_n^{-H}$; see [4]. As in the matrix sequence of increasing order $\{D_n\}_{n=1}^{\infty}$ each D_n is principal submatrix of the following D_{n+1}, we can associate the infinite matrix D with M. The upper Hessenberg matrix D gives a large recurrence relation for generating the monic orthogonal sequence $\{P_n(z)\}_{n=0}^{\infty}$, where:

$$P_n(z) = |I_n z - D_n|. \tag{4}$$

Hence, we can obtain M, D, and the monic orthogonal polynomial sequence no associated with measures. Nevertheless it has uniquely algebraic interest. In the case of the Jacobi matrix, $D = J$, it allows us the three-term recurrence relation.

2.1 Orthogonal polynomials on the real line

For the orthogonal polynomials sequences on the real line the matrix $D = J$ is the well-known Jacobi matrix,

$$J = \begin{pmatrix} b_0 & a_1 & 0 & \cdots \\ a_1 & b_1 & a_2 & \cdots \\ 0 & a_2 & b_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

with $b_i \in \mathbb{R}$ y $a_i > 0$. The Jacobi matrices J have associated a well-known three-term recurrence relation that generates the orthogonal polynomial sequence.

We characterize the matrices U, V, and T so that $J = (UV + T)^{-1}$. Note that the matrix J is generated by two numerical sequences $(a_1, a_2, \cdots, a_n, \cdots)$, $(b_0, b_1, \cdots, b_n, \cdots)$. Because J is symmetric, the real matrix $UV + T$ must also be symmetric. Then trivially,

$$t_{ij} = u_j v_i - u_i v_j. \tag{5}$$

Therefore, the orthogonal polynomial sequences on the real line are given by the sequences $(u_1, u_2, u_3, ...)$ and $(v_1, v_2, v_3, ...)$, because the matrix T satisfies (5).

2.2 Relations between the finite sequence from J_n and B_n.

Proposition 1 *Let J_n a real symmetric tridiagonal matrix generated by $(b_0, b_1, ...b_{n-1})$ and $(a_1, a_2, ..., a_{n-1})$, with $a_i > 0$. Then its inverse matrix is determined by the vectors $U = (u_1, u_2, ..., u_n)$ and $V = (v_1, v_2, ..., v_n)$ generated by the recurrence relations*

$$\begin{cases} v_1 = 1, \quad v_2 = \dfrac{1 - b_0 v_1}{a_1}, \\ v_{j+1} = \dfrac{1 - a_{j-1}v_{j-1} - b_{j-1}v_j}{a_j}, \\ j = 2, 3, ..., n-1. \end{cases} \qquad \begin{cases} u_n = \dfrac{(-1)^{n+1}[a_1 a_2 \cdots a_{n-1}]}{|J_n|}, , \\ u_{n-1} = \dfrac{(-1)^n [a_1 a_2 \cdots a_{n-2}] b_{n-1}}{|J_n|}, \\ u_{i-1} = \dfrac{-b_{i-1}u_i - a_i u_{i+1}}{a_{i-1}}, \\ i = n-1, ..., 3, 2. \end{cases} \tag{6}$$

and the matrix $T_n = (t_{ij})$ given by (5).

Proof. We take $v_1 = 1.$. The product of the first row of J_n by the first column of B_n gives:

$$b_0 u_1 v_1 + a_1 u_1 u_2 = 1 \quad \Rightarrow \quad v_2 = \frac{1 - b_0 v_1}{a_1}.$$

The product of the j-th row of J_n by the j-th column of B_n, with $u_j \neq 0$, and $a_j \neq 0$, gives:

$$a_{j-1}u_j v_{j-1} + b_{j-1}u_j v_j + a_j u_j v_{j+1} = 1 \quad \Rightarrow \quad v_{j+1} = \frac{1 - a_{j-1}v_{j-1} - b_{j-1}v_j}{a_j},$$

Hence, the entries v_j from V can recursively be computed from v_1 and v_2.

Solving u_n from (2), and taking into consideration that $v_1 = 1$ and $b_{i,i+1} - u_i v_{i+1} = t_{i+1,i}$, we have

$$u_n = \frac{(-1)^{n+1}[a_1 a_2 \cdots a_{n-1}]}{|J_n|}.$$

The product of the first row of B_n by the last column of J_n gives:

$$u_{n-1}a_{n-1} + u_n b_{n-1} = 0 \quad \Rightarrow \quad u_{n-1} = \frac{-b_{n-1}u_n}{a_{n-1}} = \frac{(-1)^n [a_1 a_2 \cdots a_{n-2}] b_{n-1}}{|J_n|}.$$

The product of the first row of B_n by the i-th column of J_n gives:

$$u_{i-1}a_{i-1} + u_i b_{i-1} + u_{i+1}a_i = 0 \quad \Rightarrow \quad u_{i-1} = \frac{-b_{i-1}u_i - a_i u_{i+1}}{a_{i-1}}.$$

Hence, the entries u_i of U are recursively computed from u_n and u_{n-1}. The matrix T_n, with entries given by (5), is required so that B_n be a symmetric matrix. ■

Proposition 2 *Given $B_n = (b_{ij})_{i,j=1}^n$ with the structure $UV + T$, and the t_{ij} satisfying* (5)*. Then its inverse matrix, symmetric and tridiagonal, J_n is defined by the sequences $\{b_i\}_{i,j=1}^n$ and $\{a_i\}_{i,j=1}^n$, $a_i > 0$. Moreover, such sequences satisfy the recursive relations:*

$$a_i = \frac{1}{u_{i+1}v_i - u_i v_{i+1}}, \qquad b_i = \frac{1 - a_i u_{i+1} - a_{i+1}u_{i+2}v_{i+1}}{u_{i+1}v_{i+1}}. \tag{7}$$

Proof. The entries on the subdiagonal from J_n are the reciprocal of the correspondent entries on the superdiagonal from the matrix T, and its value is given by (5). Hence:

$$a_i = h_{i+1,1} = \frac{1}{t_{i,i+1}} = \frac{1}{u_{i+1}v_i - u_i v_{i+1}}.$$

Known the a_i, we can obtain the b_i considering the product of the i-th row of J_n by the i-th column of B_n:

$$a_i u_{i+1} + b_i u_{i+1}v_{i+1} + a_{i+1}u_{i+2}v_{i+1} = 1 \quad \Rightarrow \quad b_i = \frac{1 - a_i u_{i+1} - a_{i+1}u_{i+2}v_{i+1}}{u_{i+1}v_{i+1}}.$$

■

3 Consistency of the finite Hessenberg matrices

We consider now finite nonsingular $n \times n$ matrices B with the same structure $B = UV + T$, i.e. a rank one perturbation of a tridiagonal matrix with null diagonal and full positive superdiagonal, with the conditions $u_k \neq 0$, $\forall k = 1, 2, ..., n$ and $v_1 \neq 0$. The principal sections $B_1, B_2, ..., B_n$ of B, have also the same structure $UV + T$. Then, their inverses are also unreduced upper Hessenberg matrices with positive subdiagonal.

The natural question about the construction of such sections of the Hessenberg matrix H is related with the consistency, if H_k is a principal submatrix of H_{k+1}. This is no true because the last column of H_k is different than the correspondent column in H_{k+1}. Nevertheless, it is true that

$$(H_k)_{k-1} = (H_{k+1})_{k-1}.$$

That is, when deleting the last column and row of H_k, and the two last columns and rows of H_{k+1}, the resulting matrices are the same. In the following theorem is shown that this matrix sequence $\{H_k\}$ is consistent. The matrix H can be built by sections.

Theorem 1 *Given the $n \times n$ matrix B with the structure $UV + T$, $u_k \neq 0$, $\forall k = 1, 2, ..., n$ and $v_1 \neq 0$. When the H_k, principal sections of the matrix B^{-1}, and their inverses are considered, the sequence of the inverses $\{H_k\}$ is consistent. That is, each matrix contains the entries of the previous one, with the exception of its last row and column, and satisfying $(H_k)_{k-1} = (H_{k+1})_{k-1}$, $\forall k = 2, 3, ..., n-1$.*

Proof. Let B_k be the k-th section of B, and H_k the inverse matrix of B_k. In order to demonstrate the consistency of H_k, we assume the following block partition for the B_k,

$$\left(\begin{array}{cccccc|c}
u_1v_1 & u_1v_2+t_{12} & u_1v_3+t_{13} & u_1v_4+t_{14} & \cdots & u_1v_{k-1}+t_{1,k-1} & u_1v_k+t_{1k} \\
u_2v_1 & u_2v_2 & u_2v_3+t_{23} & u_2v_4+t_{24} & \cdots & u_2v_{k-1}+t_{2,k-1} & u_2v_k+t_{2k} \\
u_3v_1 & u_3v_2 & u_3v_3 & u_3v_4+t_{34} & \cdots & u_3v_{k-1}+t_{3,k-1} & u_3v_k+t_{3k} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
u_{k-1}v_1 & u_{k-1}v_2 & u_{k-1}v_3 & u_{k-1}v_4 & \cdots & u_{k-1}v_{k-1} & u_{k-1}v_k+t_{k-1,k} \\
\hline
u_kv_1 & u_kv_2 & u_kv_3 & u_kv_4 & \cdots & u_kv_{k-1} & u_kv_k
\end{array}\right)$$

and for

$$H_k=\left(\begin{array}{cccccc|c}
h_{11} & h_{12} & h_{13} & h_{14} & \cdots & h_{1,k-1} & h_{ik} \\
h_{21} & h_{22} & h_{23} & h_{24} & \cdots & h_{2,k-1} & h_{2k} \\
0 & h_{32} & h_{33} & h_{34} & \cdots & h_{3,k-1} & h_{3k} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & h_{k-1,k-1} & h_{k-1,k} \\
\hline
0 & 0 & 0 & 0 & \cdots & h_{k,k-1} & h_{kk}
\end{array}\right).$$

The product B_kH_k accomplishes:

$$B_kH_k=\left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22}\end{array}\right)\left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & H_{22}\end{array}\right)=\left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & 1\end{array}\right),$$

and we derive:

$$\begin{aligned}
B_{11}H_{11}+B_{12}H_{21} &= B_{11}H_{11}+\begin{pmatrix} u_1v_k+t_{1k} \\ u_2v_k+t_{2k} \\ u_3v_k+t_{3k} \\ \vdots \\ u_{k-1}v_k+t_{k-1,k}\end{pmatrix}\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & h_{k,k-1}\end{pmatrix} \\
&= B_{11}H_{11}+\begin{pmatrix} 0 & 0 & \cdots & 0 & (u_1v_k+t_{1k})h_{k,k-1} \\ 0 & 0 & \cdots & 0 & (u_2v_k+t_{2k})h_{k,k-1} \\ 0 & 0 & \cdots & 0 & (u_3v_k+t_{3k})h_{k,k-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & (u_{k-1}v_k+t_{k-1,k})h_{k,k-1}\end{pmatrix}=I_{k-1}.
\end{aligned}$$

Taking the $(k-2)$-th section from these matrices of order $k-1$, in order to avoid the nonzero column, we have:

$$(B_{11}H_{11}+B_{12}H_{21})_{k-2}=(B_{11}H_{11})_{k-2}=I_{k-2}.$$

Since the matrix H_{11} is upper Hessenberg:

$$(B_{11})_{k-2}(H_{11})_{k-2}=(B_{11}H_{11})_{k-2}=I_{k-2},\quad k=3,4,...,n.$$

From this result, we finally obtain, $(H_k)_{k-1} = ((B_k)^{-1})_{k-1} = (H_n)_{k-1}$. ■

This result aim us to define a matrix sequence $\{H_k\}_{k=1}^{n}$ with an increasing order, where each matrix is a principal submatrix of the following matrices in the sequence, i.e. allows us to define the matrix $H_n = (h_{ij})_{i,j=1}^{n}$ associated to the matrix B. This possibility could be many applications in the case of infinite matrices.

Example 2 *Given the upper Hessenberg matrix of order* 6*:*

$$D = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 & 3 \\ 0 & 1 & 2 & 1 & -1 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$$

with inverse matrix:

$$B = \begin{pmatrix} 5/6 & -2/3 & 1/2 & -1/3 & 1/6 & 1/2 \\ -1/2 & 1 & -1/2 & 0 & 1/2 & -3/2 \\ 2/3 & -4/3 & 2 & -5/3 & 4/3 & 1 \\ -1/2 & 1 & -3/2 & 2 & -3/2 & -1/2 \\ 1/3 & -2/3 & 1 & -4/3 & 5/3 & 0 \\ -1/6 & 1/3 & -1/2 & 2/3 & -5/6 & 1/2 \end{pmatrix}$$

Taking the sections B_2, B_3, B_4 y B_5 of the matrix B, and inverting such matrices, we have:

$$D_2 = \begin{pmatrix} 2 & 4/3 \\ 1 & 5/3 \end{pmatrix}, \; D_3 = \begin{pmatrix} 2 & 1 & -1/4 \\ 1 & 2 & 1/4 \\ 0 & 1 & 3/4 \end{pmatrix}$$

$$D_4 = \begin{pmatrix} 2 & 1 & 0 & 1/3 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 2 & 5/3 \\ 0 & 0 & 1 & 4/3 \end{pmatrix}, \; D_5 = \begin{pmatrix} 2 & 1 & 0 & 0 & -1/2 \\ 1 & 2 & 1 & 0 & -3/2 \\ 0 & 1 & 2 & 1 & -1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 3/2 \end{pmatrix}.$$

We can observe that $(D_3)_1 = (D_2)_1$, $(D_4)_2 = (D_3)_2$, $(D_5)_3 = (D_4)_3$. It is the recursive form to generate the matrix D.

4 Matrices U, V and T in terms of orthogonal polynomials

Let M be an infinite HDP matrix and let D be the associated Hessenberg matrix. The $\{P_n(z)\}$ are the monic polynomials, the $\{p_n(z)\}$ the normalized polynomials, and $\{P_{n-j}^{(j)}(z)\}$,

$n > j$, $P_0^{(j)}(z) = 1$, be the associated monic polynomials, defined by $P_{n-j}^{(j)}(z) = |I_{n-j}z - D_{n-j}^{(j)}|$, where $D_{n-j}^{(j)}$ is the matrix of order $n-j$. This matrix results when deleting the j first rows and columns of the matrix D_n. Also, let $\{p_{n-j}^{(j)}(z)\}$ be the normalized associated polynomials, see [7], where the normalization considered for the monic polynomials is given by formula $p_{n-j}^{(j)}(z) = \frac{P_{n-j}^{(j)}(z)}{[d_{j+2,j+1}\cdots d_{n+1,n}]}$, $\quad n > j, \quad p_0^{(j)} = 1$. For $j = 0$, this normalization is also adequate for the polynomials $p_n(z)$. The n-th resolvent matrix is defined as $(I_nz - D_n)^{-1}$, $\forall z \in \mathbb{C}\backslash\{z : p_n(z) = 0\}$.

Theorem 2 *The finite resolvent matrices associated to matrix D are given by*

$$(I_nz - D_n)^{-1}[i,j] = \begin{cases} \dfrac{1}{d_{i+1,i}} \dfrac{p_{j-1}(z)p_{n-i}^{(i)}(z)}{p_n(z)}, & \textit{if } j \le i, \\ \dfrac{1}{d_{i+1,i}} \left[\dfrac{p_{j-1}(z)p_{n-i}^{(i)}(z)}{p_n(z)} - p_{j-i-1}^{(i)}(z)\right], & \textit{if } j > i. \end{cases} \tag{8}$$

Proof. We must multiply $I_nz - D_n$ by $(I_nz - D_n)^{-1}$, and we have I_n. For this task, it is sufficient multiply the i-th row of $I_nz - D_n$ by the i-th column of $(I_nz - D_n)^{-1}$. Using formulas of the large recurrence relation satisfied for the associated polynomials, we obtain that this product is 1. Also we multiply the i-th row of first matrix by the j-th column, $i \ne j$, of the second matrix. In a similar way, we obtain that this product is 0. ■

4.1 A method to compute the inverses of finite Hessenberg matrices

Taking $z = 0$ in expression (8), with i and j determined, the left side is a number with a minus sign. Multiplying by (-1) in both sides, we have

$$D_n^{-1}[i,j] = \begin{cases} \dfrac{-1}{d_{i+1,i}} \dfrac{p_{j-1}(0)p_{n-i}^{(i)}(0)}{p_n(0)}, & \text{if } j \le i, \\ \dfrac{-1}{d_{i+1,i}} \dfrac{p_{j-1}(0)p_{n-i}^{(i)}(0)}{p_n(0)} + \dfrac{1}{d_{i+1,i}} p_{j-i-1}^{(i)}(0), & \text{if } j > i. \end{cases} \tag{9}$$

As a consequence, if the polynomials are known, we can compute the inverse of the matrix D_n with this method. It is an alternative to the method given in Lemma 1. Finally, we give a numerical example of computation of the inverse matrix from the polynomial sequence.

Example 3 *Let be the tridiagonal matrix* $J_n = \begin{pmatrix} 2 & 1 & 0 & & \\ 1 & 2 & 1 & & \\ & 1 & 2 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 2 \end{pmatrix}$, *where, as* $d_{i+1,i} = 1$, *the norm of monic polynomials is* 1 *and* $p_n(z) = P_n(z)$. *The determinants* $|J_n| = n+1$ *are* *obtained by induction. The polynomials for $z = 0$ are $p_n(0) = (-1)^n(n+1)$. The associated polynomials are the same than the general polynomials because the matrix is Toeplitz. Then, the inverse matrix results as $J_n^{-1} = UV + T$, with*

$$U = \begin{pmatrix} \frac{n}{n+1} \\ -\frac{n-1}{n+1} \\ \frac{n-2}{n+1} \\ \vdots \\ (-1)^{n-1}\frac{1}{n+1} \end{pmatrix}, \quad V = \begin{pmatrix} 1 & -2 & 3 & -4 & \cdots & (-1)^{n-1}n \end{pmatrix},$$

and the matrix $T = \begin{pmatrix} 0 & 1 & -2 & 3 & -4 & \cdots & (-1)^{n-1}(n-2) \\ 0 & 0 & 1 & -2 & 3 & \cdots & (-1)^{n-2}(n-3) \\ 0 & 0 & 0 & 1 & -2 & \cdots & (-1)^{n-3}(n-4) \\ 0 & 0 & 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$.

Note that, when n tends to ∞, the vector U will be $U = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & \cdots \end{pmatrix}^T$, and the inverse matrix obtained, inverse of the infinite matrix J, is

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

as it is easy to verify doing the products $BJ = JB = I$.

5 Conclusions

A new method to generate Hessenberg matrices with positive subdiagonal has been proposed. That is, a method to obtain orthogonal polynomial sequences. We have studied the conditions on U, V, and T in order to have an associated symmetric tridiagonal Jacobi matrix D. Basically, T must verify $t_{ij} = u_j v_i - u_i v_j$. The conditions for matrix D will be an upper Hessenberg matrix, no Jacobi tridiagonal, are $t_{ij} \neq u_j v_i - u_i v_j$. The determination of conditions on U, V, and T that allow us know when the associated matrix D be a solution of a moment problem, i.e. the study of subnormality of matrix D, remains as an open problem.

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

[4] C. Escribano, A. Giraldo, M. A. Sastre, E. Torrano, *Hessenberg matrix for sums of Hermitian positive definite matrices and weighted shifts*, J. Comp. Appl. Math. **236** (2011) 98–106.

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

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

[7] G. Szegö, *Orthogonal Polynomials*, fourth ed., Colloq. Pub., vol. 23, Amer. Math. Soc., Rhode Island, USA, 1975

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

ParallDroid: A Framework for Parallelism in Android^{TM}

A. Acosta[1] **and F. Almeida**[1]

[1] *Dept. Estadística, I.O. y Computación, La Laguna University, Spain*

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

Abstract

The advent of emergent SoCs and MPSocs opens a new era on the small mobile devices (Smartphones, Tablets, ...) in terms of computing capabilities and applications to be addressed. The efficient use of such devices, including the parallel power, is still a challenge for general purpose programmers due to the very high learning curve demanding very specific knowledge of the devices. While some efforts are currently being made, mainly in the scientific scope, the scenario is still quite far from being the desirable for non-scientific applications where very few of them take advantage of the parallel capabilities of the devices. We propose Paralldroid (Framework for Parallelism in Android), a parallel development framework oriented to general purpose programmers for standard mobile devices. Paralldroid presents a programming model that unifies the differents programming models of Android. The user just implements a Java applications and introduces a set of Paralldroid annotations in the sections of code to be optimized. The Paralldroid system automatically generates the native C, OpenCL or Renderscript code for the annotated section. The ParallDroid transformation model involves source-to-source transformations and skeletal programming.

Key words: SoC, android, source to source translator.

1 Introduction

SoCs (Systems on Chip [1]) have been the enabling technology behind the evolution of many of todays ubiquitous technologies, such as Internet, mobile wireless technology, and high definition television. The information technology age, in turn, has fuelled a global communications revolution. With the rise of communications with mobile devices, more computing power has been put in such systems. The technologies available in desktop computers are now implemented in embedded and mobile devices. We find new processors with multicore architectures and GPUs developed for this market like the Nvidia Tegra [2] with two and five ARM cores and a low power GPU, and the OMAP^{TM}5 [3] platform from Texas Instruments that also goes in the same direction.

On the other hand, software frameworks have been developed to support the building of software for such devices. The main actors in this software market have their own platforms: Android [4] from Google, iOS [5] from Apple and Windows phone [6] from Microsoft are contenders in the smartphone market.

Conceptually, from the architectural perspective, the model can be viewed as a traditional heterogeneous CPU/GPU with a unified memory architecture, where memory is shared between the CPU and GPU and acts as a high bandwidth communication channel. In the non-unified memory architectures, it was common to have only a subset of the actual memory addressable by the GPU.

Under this scenario, we find a strong divorce among traditional mobile software developers and parallel programmers, the first tend to use high level frameworks like Eclipse for the development of Java programs, without any knowledge of parallel programming (Android: Eclipse + Java, Windows: Visual Studio + C#, IOS: XCode + Objective C), and the latter that use to work on Linux, doing their programs directly in OpenCL closer to the metal. The former take the advantage of the high level expressiveness while the latter assume the challenge of high performance programming. The work developed in this paper tries to bring these to worlds.

We propose the ParallDroid system, a development framework that allows for the automatic development of Renderscript, native C or OpenCL code for mobile devices (Smartphones, Tablets, ...). The developer fills and annotate the code that wants to execute in parallel. Annotations used are an extension of OpenMP [7]. ParallDroid uses the information provided in this annotations to generate a new parallel program that incorporates the code sections to run over mobile device.

The advantages of this approach are well known:

- Increased use of the parallel devices by non-expert users
- Rapid inclusion of emerging technology into their systems
- Delivery of new applications due to the rapid development time
- Unifies the differents programming models of Android

To validate the performance of the code generated by our framework, we consider five different applications, transform a image to grayscale, convolve 3x3 and 5x5, levels and a general convolve implementation. In all cases, we implemented five versions of code, the ad-hoc version from a Java developer, ad-hoc native C implementation, ad-hoc Renderscript implementation, and the versions generated by paralldroid, generated native C code and generated renderscript code. We execute these problems over two SoCs devices running Android, a Samsung Galaxy SIII and an Asus Transformer Prime TF201.

The computational experience performed on five different problems prove that the results are quite promising. The computational experience proved that ParallDroid offers good performances with a very low cost of development where the parallelism is hidden to the sequential developer. As expected, the renderscript versions improve to the Java implementation so, ParallDroid is an useful tool for the automatic generation of the renderscript code. ParallDroid also contributes to increase the productivity in the parallel developments due to the low effort required.

Acknowledgements

This work has been supported by the EC (FEDER) and the Spanish MEC with the I+D+I contract number: TIN2011-24598

References

[1] SoCC: IEEE International System–on–Chip Conference. http://www.ieee-socc.org/ (September 2012)

[2] NVIDIA: NVIDIA Tegra mobile processors: Tegra2, Tegra 3 and Tegra 4. http://www.nvidia.com/object/tegra-superchip.html

[3] Instruments, T.: OMAP^{TM}Mobile Processors : OMAP^{TM}5 platform. http://www.ti.com/omap5

[4] Google: Android mobile platform. http://www.android.com

[5] Apple: iOS: Apple mobile operating system. http://www.apple.com/ios

[6] Microsoft: Windows Phone: Microsoft mobile operating system. http://www.microsoft.com/windowsphone

[7] OpenMP: Openmp: Api specification for parallel programming. http://openmp.org/wp/

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

"kâi lêuat òk" is everything: 26, 27, 66

Maíra Aguiar[1] **and Nico Stollenwerk**[1]

[1] *Centro de Matemática e Aplicações Fundamentais, Lisbon University*

emails: maira@ptmat.fc.ul.pt, nico@eptmat.fc.ul

Abstract

Dengue fever dynamics are well known to be particularly complex with large fluctuations of disease incidence. Mathematical models describing the transmission of dengue viruses need to be parametrized on data referring to incidence of disease in order to be used as a predictive tools to evaluate the introduction of intervention strategies like vector control and vaccination. For Thailand, two different sources of the long-term empirical dengue data (1980-2012) are available and consists of monthly incidences of hospital admission cases, the hard copy (HC) data from 1980-2005 and the electronic files (EF) from 2003 to present, where the clinical classification of the disease are notified using separate files. By gathering, cross-checking and analyzing the overlapping epidemiological years of the data (2003-2005) we observed a considerable underestimation of dengue cases in Thailand during the transition HC-EF, which affects considerably the model development, interpretation and its correct application.

Key words: Dengue fever, data analysis, modeling

1 Introduction

It is estimated that every year there are $70 - 500$ million dengue infections, 36 million cases of dengue fever (DF) and 2.1 million cases of dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), with more than 20.000 deaths per year [2, 1]. In many countries in Asia and South America DF and DHF/DSS has become a substantial public health concern leading to serious social-economic costs. Over the past 50 years, the number of infected patients (DF/DHF/DSS) has been rising steadily. 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 and besides the relatively slow progress in the development of the dengue virus vaccine, a tetravalent live attenuated vaccine and a live chimeric virus vaccine are presently going through Phase III trials [3, 4].

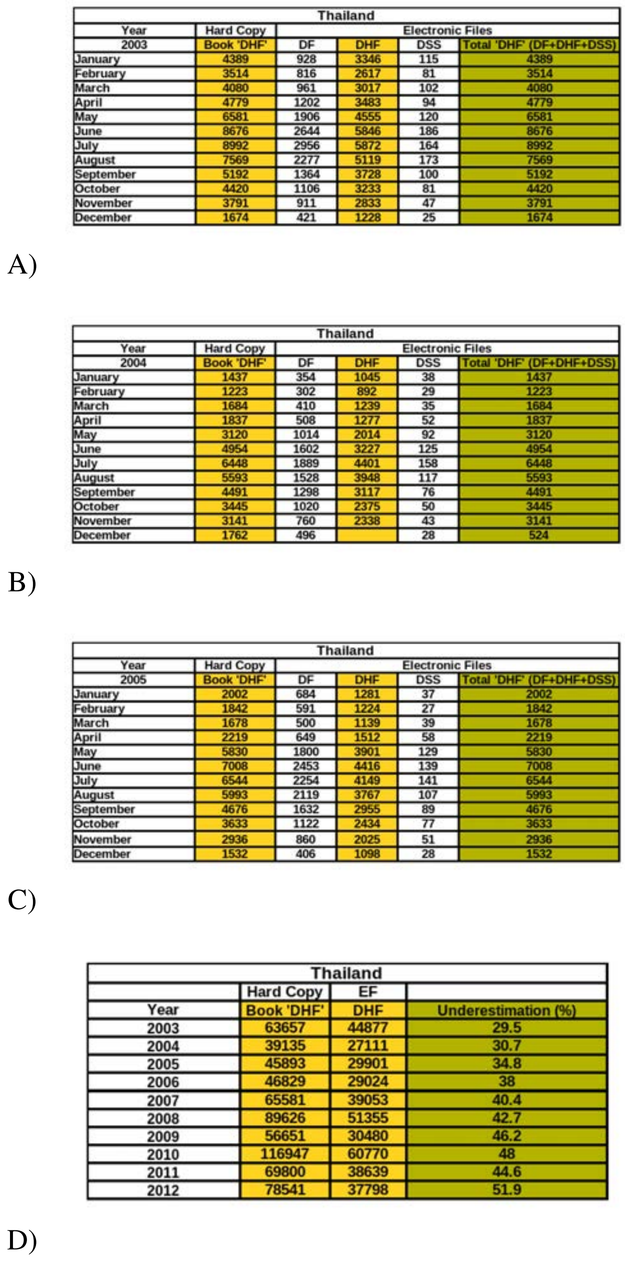
The Thai word used to describe dengue illness refers to fever with blood leakage (kâi lêuat òk, the English pronunciation of the Thai language), which was later on possibly written and translated according to each one of the possible clinical manifestation of the disease, as it follows: (I) kâi lêuat òk deeng gèe or code 26, to describe DHF, however it is unusual to be found in the Thai documents, where the simplification kâi lêuat òk is mostly used. The DHF cases may evolve towards a collection of symptoms with hemorrhagic fever leading to shock or DSS, in Thai (II) glùm aa-gaan kâi lêuat òk chòk or code 27. The classic dengue or DF, a non-fatal form of illness classically defined as a fever caused by a dengue virus (which is rarely associated with blood leakage), is traditionally written as kâi lêuat òk but could be eventually written as (III) kâi deeng gèe or code 66.

Dengue dynamics is well known to be particularly complex with large fluctuations of disease incidences. The long-term incidence dengue data in Thailand have been continually used, among theoretical epidemiologists and modelers, to describe the dynamics of dengue epidemics. Up to now, 33 years of dengue illness incidence data are available and have been continually used, as for example in [5, 6, 18, 8, 9], to parametrize mathematical models as a tool to predict and control the disease.

From 1980 to 2005 the aggregated data, presented as “DHF-Total”, have been publicly distributed as a hard copy (HC) book format through the Bureau of Epidemiology Annual Epidemiological Surveillance Report [21]. Starting in 2003, the disease was notified using separate electronic files (EF) for each one of the possible clinical classification. By gathering, cross-checking and analyzing the overleaping data (2003-2005) it was observed that the sum of all clinical cases of the disease (DF+DHF+DSS) gives rise to the DHF-Total incidence data in Thailand. The English translation of the Thai documents still causes confusion when interpreting the data, and for modeling purposes it leads to a significant underestimation of the real number of dengue cases in Thailand. From 2003 on, only the clinical classification of DHF cases have been counted whereas the DF and DSS cases were neglected. Such data cross-checking was carried out for all Provinces in Thailand showing the same results.

In Fig. 1A we present the official monthly incidence of dengue illness in Thailand for both sources of the 2003 overlapping data, the HC [21] and the EF separate files for each disease clinical classification. We list the given numbers for DHF-Total, DF, DHF and DSS respectively and observe that the number of cases classified clinically as DHF differs considerably from the numbers presented as DHF-Total in [21]. When taking into account also the number of cases for DF and DSS we clearly see that the final numbers matches the original data collection of DHF-Total cases. For the epidemiological years of 2004 and 2005, the results were exactly the same (see Fig. 1B-C, and the underestimation of dengue cases is increasing rapidly as shown in Fig. 1D.

This underestimation is strongly associated with the English translation and misinterpretation of the Thai documents, generating an inexact continuous time series to be used for modeling purposes, affecting considerably the model development, interpretation and its correct application.



Thailand					
Year	Hard Copy	Electronic Files			
2003	Book 'DHF'	DF	DHF	DSS	Total 'DHF' (DF+DHF+DSS)
January	4389	928	3346	115	4389
February	3514	816	2617	81	3514
March	4080	961	3017	102	4080
April	4779	1202	3483	94	4779
May	6581	1906	4555	120	6581
June	8676	2644	5846	186	8676
July	8992	2956	5872	164	8992
August	7569	2277	5119	173	7569
September	5192	1364	3728	100	5192
October	4420	1106	3233	81	4420
November	3791	911	2833	47	3791
December	1674	421	1228	25	1674

A)

Thailand					
Year	Hard Copy	Electronic Files			
2004	Book 'DHF'	DF	DHF	DSS	Total 'DHF' (DF+DHF+DSS)
January	1437	354	1045	38	1437
February	1223	302	892	29	1223
March	1684	410	1239	35	1684
April	1837	508	1277	52	1837
May	3120	1014	2014	92	3120
June	4954	1602	3227	125	4954
July	6448	1889	4401	158	6448
August	5593	1528	3948	117	5593
September	4491	1298	3117	76	4491
October	3445	1020	2375	50	3445
November	3141	760	2338	43	3141
December	1762	496		28	524

B)

Thailand					
Year	Hard Copy	Electronic Files			
2005	Book 'DHF'	DF	DHF	DSS	Total 'DHF' (DF+DHF+DSS)
January	2002	684	1281	37	2002
February	1842	591	1224	27	1842
March	1678	500	1139	39	1678
April	2219	649	1512	58	2219
May	5830	1800	3901	129	5830
June	7008	2453	4416	139	7008
July	6544	2254	4149	141	6544
August	5993	2119	3767	107	5993
September	4676	1632	2955	89	4676
October	3633	1122	2434	77	3633
November	2936	860	2025	51	2936
December	1532	406	1098	28	1532

C)

Thailand			
	Hard Copy	EF	
Year	Book 'DHF'	DHF	Underestimation (%)
2003	63657	44877	29.5
2004	39135	27111	30.7
2005	45893	29901	34.8
2006	46829	29024	38
2007	65581	39053	40.4
2008	89626	51355	42.7
2009	56651	30480	46.2
2010	116947	60770	48
2011	69800	38639	44.6
2012	78541	37798	51.9

D)

Figure 1: Thailand DHF-Total data comparison between the HC book and EF for each clinical classification DF, DHF and DSS. In A) 2003 epidemiological year, in B) 2004 epidemiological year and in C) 2005 epidemiologycal year. In D) from 2003 to present, Thailand underestimation of dengue cases.

2 Compartmental models applied to dengue fever

Almost all mathematical models for infectious diseases start from the same basic premise: that the population can be subdivided into a set of distinct classes. The most commonly used framework for epidemiological systems, is still the SIR-type model, a good and simple model for many infectious diseases. The SIR epidemic model divides the population into three classes: susceptible (S), Infected (I) and Recovered (R). Multi-strain dengue models are modeled with SIR-type models where the SIR classes are labeled for the hosts that have seen the individual strains. Assuming that the transmission of the disease is contagious from person to person, the susceptibles become infected and infectious with infection rate β, are cured and become recovered with recovery rate γ. Partial or temporary or life long immunity can be assumed, depending on the biology of the pathogen After a waning immunity period α, the recovered individual can become susceptible, to a different strain again.

Retrospective dengue data and the possibility to estimate hidden states in dengue models from such data [8, 10] have been discussed, specially, primary versus secondary infection, and symptomatic versus asymptomatic cases that can be studied via the first already available models [11, 12, 8, 9, 13]. Although the multi-strain interaction leading to deterministic chaos via ADE has been described previously, e.g. [14, 15, 16], the role of temporary cross-immunity has been neglected leading to unrealistic biological parameter estimation. More recently, despite incorporation of temporary cross-immunity in rather complicated models, the possible dynamical structures were not deeply analyzed [17, 18, 19]. When including temporary cross-immunity into the ADE dengue models, a rich dynamic structure including chaos in wider and more biologically realistic parameter regions was found [11, 12, 8, 9, 13].

The basic two-strain model shows a rich variety of dynamics through bifurcations up to deterministically chaotic behavior in wider and more biologically realistic parameter regions than previously anticipated when neglecting temporary cross-immunity. The seasonal two-strain model with import of infected have shown a qualitatively good result when comparing empirical dengue data and simulation results, where patterns of the data behavior were similarly found to happen in the time series simulations (see Fig. 2A).

2.1 The two-strain model

The two-strain model with temporary cross-immunity is a 9 dimensional system where the population N is divided into ten classes. For two different strains, named strain 1 and strain 2, we label the SIR classes for the hosts that have seen the individual strains, without epidemiological asymmetry between strains, i.e. infections with strain one followed by strain two or vice versa contribute in the same way to the force of infection. The complete system of ordinary differential equations (ODEs) for the two-strain epidemiological model can be written as follows.

$$
\begin{aligned}
\dot{S} &= \mu(N-S) - \frac{\beta(t)}{N} S(I_1 + \rho\cdot N + \phi I_{21}) \\
&\quad - \frac{\beta(t)}{N} S(I_2 + \rho\cdot N + \phi I_{12}) && (1)\\
\dot{I_1} &= \frac{\beta(t)}{N} S(I_1 + \rho\cdot N + \phi I_{21}) - (\gamma+\mu) I_1 && (2)\\
\dot{I_2} &= \frac{\beta(t)}{N} S(I_2 + \rho\cdot N + \phi I_{12}) - (\gamma+\mu) I_2 && (3)\\
\dot{R_1} &= \gamma I_1 - (\alpha+\mu) R_1 && (4)\\
\dot{R_2} &= \gamma I_2 - (\alpha+\mu) R_2 && (5)\\
\dot{S_1} &= \alpha R_1 - \frac{\beta(t)}{N} S_1(I_2 + \rho\cdot N + \phi I_{12}) - \mu S_1 && (6)\\
\dot{S_2} &= \alpha R_2 - \frac{\beta(t)}{N} S_2(I_1 + \rho\cdot N + \phi I_{21}) - \mu S_2 && (7)\\
\dot{I_{12}} &= \frac{\beta(t)}{N} S_1(I_2 + \rho\cdot N + \phi I_{12}) - (\gamma+\mu) I_{12} && (8)\\
\dot{I_{21}} &= \frac{\beta(t)}{N} S_2(I_1 + \rho\cdot N + \phi I_{21}) - (\gamma+\mu) I_{21} && (9)\\
\dot{R} &= \gamma(I_{12} + I_{21}) - \mu R \quad . && (10)
\end{aligned}
$$

2.2 Numerical analysis

Mathematical models describing the transmission of dengue viruses need to be parametrized on data referring to incidence of disease. We take Chiang Mai Province, North of Thailand, as match empirical data with model simulations. The birth and death rate, recovery rate, degree of seasonality and the temporary cross-immunity rate are fixed. The infection rate and ratio of secondary infections contributing to force of infection are the parameters that may vary when matching different data sets.

In Figure 2A we match empirical DHF-Total (DF+DHF+DSS) incidence data for the Province of Chiang Mai with the two-strain model simulation. A qualitatively good result is obtained, where patterns of the irregular data occurs and is predicted by the models. The parameter values used in the simulation are listed in Table 1.

In Figure 2B we match empirical DHF-Total (DF+DHF+DSS) from 1980 to 2002 and continue with only clinical DHF incidence data with the two-strain model simulation. With a different parameter set we observe a qualitatively good result from 2003 on, however, the previous HC data can not be described. The parameter values used in the simulation are listed in Table 2.

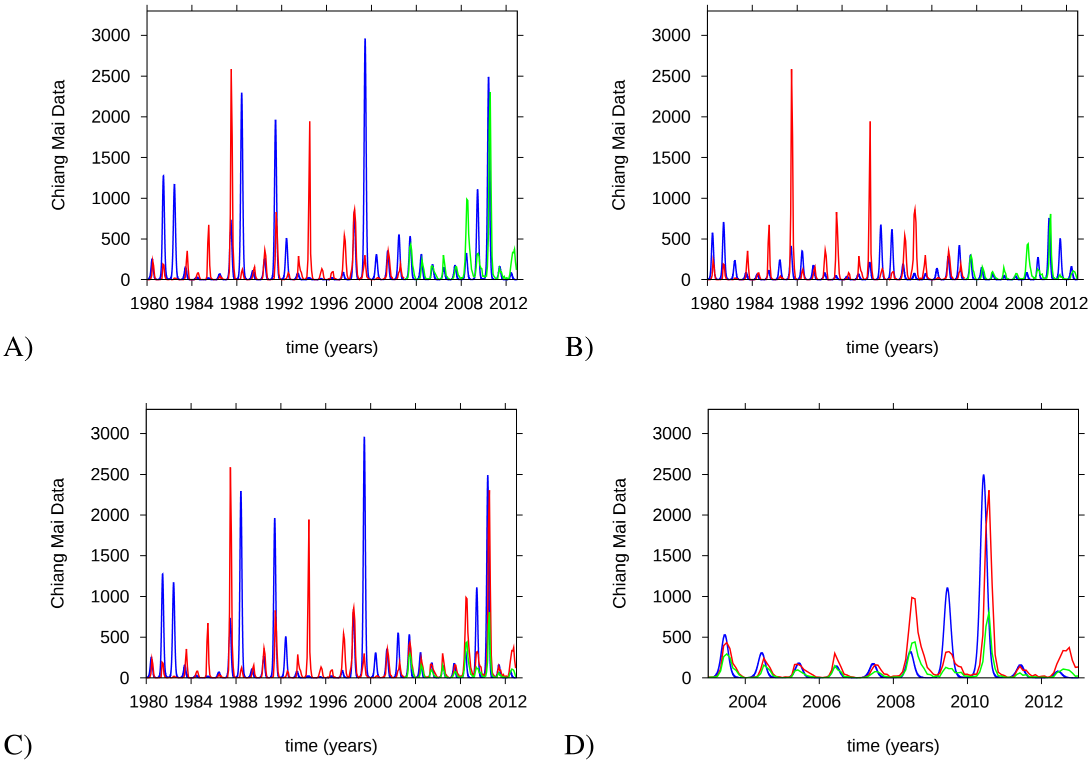


Figure 2: In A) empirical DHF-Total (DF+DHF+DSS) incidence data (in red) for the Province of Chiang Mai in the North of Thailand, with a population size $N = 1.6 \times 10^6$, are matched with simulations (in blue) for the seasonal two-strain model with import of infected. Here, the force of infection $\beta = 2\gamma$ and the secondary infection contribution to the force of infection ratio $\phi = 0.9$. The other parameter values are given in Table 1. In B) empirical DHF-Total (DF+DHF+DSS) incidence data (in red) for the Province of Chiang Mai in the North of Thailand, with a population size $N = 1.6 \times 10^6$, are matched with simulations (in blue) for the seasonal two-strain model with import of infected. Here, the force of infection $\beta = 2\gamma$ and the secondary infection contribution to the force of infection ratio $\phi = 0.9$. The other parameter values are given in Table 1. In B) we match the empirical DHF-Total (DF+DHF+DSS) from 1980 to 2002 (in red), continued with DHF clinical incidence only (in green) with the seasonal two-strain model with import of infected simulations (in blue). Here, the force of infection is considerably smaller $\beta = 1.5\gamma$ as well the secondary infection contribution to the force of infection ratio $\phi = 0.7$. The other parameter values are given in Table 2. In C-D) we present the matching of the two data sets, using parameter set ONE, with the model simulation (in blue).

Table 1: Parameter set ONE: DHF-Total (DF+DHF+DSS) data set matching

Par.	Description	Values	Ref.
N	population size	1.6×10^6	
μ	birth and death rate	$1/65y$	[22]
γ	recovery rate	$52y^{-1}$	[1]
β	infection rate	$2 \cdot \gamma$	[14, 11, 8, 9]
η	degree of seasonality	$\in [0, 0.35]$	[8, 9]
ρ	import parameter	$\in [0, 10^{-10}]$	[8, 9]
α	temporary cross-immunity rate	$2y^{-1}$	[20]
ϕ	ratio of secondary infections contributing to force of infection	0.9	[8, 9]

Table 2: Parameter set TWO: From 2003 on, DHF-only data set matching

Par.	Description	Values	Ref.
N	population size	1.6×10^6	
μ	birth and death rate	$1/65y$	[22]
γ	recovery rate	$52y^{-1}$	[1]
β	infection rate	$1.5 \cdot \gamma$	[14, 11, 8, 9]
η	degree of seasonality	$\in [0, 0.35]$	[8, 9]
ρ	import parameter	$\in [0, 10^{-10}]$	[8, 9]
α	temporary cross-immunity rate	$2y^{-1}$	[20]
ϕ	ratio of secondary infections contributing to force of infection	0.7	[8, 9]

Using the simulation obtained with parameter set ONE (see Table 1), Figures 2C-D show the data matching comparison. As for the non fixed parameters, infection rate and secondary infection contribution to the force of infection rate, we observe a considerable difference in parameter values that could lead to non efficient intervention on disease control and to a wrong Public Health intervention measures decision, depending on which data set we are using.

3 Conclusions

The two-strain model model has the minimal degree of complexity to generate both primary and secondary infection cases, and it allows to derive dengue hemorrhagic fever cases which constitute the most widely available data on dengue epidemiology. The simplicity of our model allows rigorous inference of both parameters from empirical time series.

Much of the dengue data available to theoretical epidemiologists consists of time series tracking the evolution of a subset of state variables of an underlying dynamical system through a surveillance system. For Thailand, two different sources of the long-term empirical dengue data (1980-2012) are available and consists of monthly incidences of hospital admission cases, the HC from 1980-2005 and the EF from 2003 to present. By gathering, cross-checking and analyzing the overlapping epidemiological years of the data (2003-2005) we observed a considerable underestimation of dengue cases in Thailand during the transition HC-EF files, from 2003 on, where only the clinical classification of DHF cases have been counted. The exact continuous time series is obtained when considering all admission cases, DF+DHF+DSS, however the underestimation of cases are increasing, probably a consequence of the English interpretation of the Thai documents.

As for modeling purposes two different parameter sets were obtained depending on the data set was used. Based on this observation, we conclude that the underestimation of the real number of dengue cases in Thailand will affect considerably the model development, interpretation and its correct application. Being able to describe and predict the future expected outbreaks of dengue in the absence of human interventions is the major goal if one wants to understand the effects of considered control measures. This is especially important with the upcoming perspective of a dengue virus vaccine requiring a high quality data and knowledge on how to best implement the vaccination program and use the candidate vaccine at the population-level when it will be accessible. For any interpretation based on the the long-term empirical incidence dengue data, the data aggregation is essential to improve model development, interpretation and its correct application.

Acknowledgements

Dengue surveillance data were provided by the Bureau of Epidemiology, Department off Disease Control, Ministry of Public Health, Thailand. This work has been supported by the European Union under FP7 in the DENFREE project and the Portuguese FCT project PTDC/MAT/115168/2009.

References

[1] WORLD HEALTH ORGANIZATION, *Dengue and severe dengue, Fact sheet 117* (2012). Retrieved from $http : //www.who.int/mediacentre/factsheets/fs117/en/$

[2] DENGUE VACCINE INITIATIVE (DVI), *Dengue Overview* (2013). Retrieved from $http : //www.denguevaccines.org/dengue - overview - 0$

[3] WORLD HEALTH ORGANIZATION, *Initiative for Vaccine Research* (2011). Retrieved from $http : //www.who.int/vaccine\_research/diseases/dengue/dengue_vaccines/en/index.html$

[4] SCHMITZ, J., ROEHRIG, J., BARRETT, A., HOMBACH, J., *Next generation dengue vaccines: A review of candidates in preclinical*, Vaccine **29** (2011) 7276–7284.

[5] DEREK A.T. CUMMINGS, RAFAEL A. IRIZARRY, NORDEN E. HUANG, TIMOTHY P. ENDY, ANANDA NISALAK, KUMNUAN UNGCHUSAK & DONALD S. BURKE, *Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand*, Nature **427** (2005) 334–347.

[6] CAZELLES B, CHAVEZ M, MCMICHAEL AJ, HALES S., *Nonstationary influence of El Ninõ on the synchronous dengue epidemics in Thailand*, PLoS Med. (2005) 2(4):e106.

[7] Y. NAGAO & K KOELLE, *Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever*, Proc. Natl. Acad. Sci. **105** (2008) 2238–2243.

[8] AGUIAR, M., BALLESTEROS, S., KOOI, B.W., & STOLLENWERK, N., *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** (2011) 181–196.

[9] AGUIAR, M., KOOI, W. B., ROCHA, F., GHAFFARI, P. AND STOLLENWERK, N., *How much complexity is needed to describe the fluctuations observed in dengue hemorrhagic fever incidence data?* Ecological Complexity (2012) Published online at: $http://dx.doi.org/10.1016/j.ecocom.2012.09.001$

[10] STOLLENWERK, N., AGUIAR, M., BALLESTEROS, S., BOTO, J., KOOI, W. B., & MATEUS, L., *Dynamic noise, chaos and parameter estimation in population biology*, Interface, Focus **2** (2012) 156–169.

[11] AGUIAR, M., KOOI, B., & STOLLENWERK, N., *Epidemiology of dengue fever: A model with temporary cross-immunity and possible secondary infection shows bifurcations and chaotic behaviour in wide parameter regions*, Math. Model. Nat. Phenom. **3** (2008) 48–70.

[12] AGUIAR, M., STOLLENWERK, N., & KOOI, B., *Torus bifurcations, isolas and chaotic attractors in a simple dengue fever model with ADE and temporary cross immunity*, Intern. Journal of Computer Mathematics **86** (2009) 1867–77.

[13] MAIRA AGUIAR (2012). *Rich dynamics in multi-strain models: non-linear dynamics and deterministic chaos in dengue fever epidemiology*, *PhD Thesis*, ISBN 978-989-20-2882-8, Edited by Maíra Aguiar, Portugal.

[14] FERGUSON, N., ANDERSON, R. AND GUPTA, S., *The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens*, Proc. Natl. Acad. Sci. USA 96 (1999) 790–94.

[15] SCHWARTZ, I. B., ET AL., *Chaotic desynchronization of multi-strain diseases*, Physical Review **E 72** (2005) 066201–6.

[16] BILLINGS, L., ET AL., *CInstabilities in multiserotype disease models with antibody-dependent enhancement*, Journal of Theoretical Biology 246 (2007) 18–27.

[17] Wearing, H.J. & Rohani, P. (2006). Ecological and immunological determinants of dengue epidemics *Proc. Natl. Acad. Sci. USA* , 103, 11802–11807.

[18] NAGAO, Y. & KOELLE, K., *Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever*, Proc. Natl. Acad. Sci. USA 105 (2008) 2238–2243.

[19] RECKER, M. ET AL., *Immunological serotype interactions and their effect on the epidemiological pattern of dengue*, Proc. R. Soc. B. 276 (2009) 2541–2548.

[20] S. MATHEUS, X. DEPARIS, B. LABEAU, J. LELARGE, J. MORVAN, P. DUSSART, *Discrimination between Primary and Secondary Dengue Virus Infection by an Immunoglobulin G Avidity Test Using a Single Acute-Phase Serum Sample.* Journal of Clinical Microbiology **45** (2005) 2793–97.

[21] BUREAU OF EPIDEMIOLOGY, MINISTRY OF PUBLIC HEALTH OF THAILAND, *Annual epidemiological surveillance report, Thailand* (2003, 2004, 2005) ISSN 0857-6521.

[22] WORLD POPULATION PROSPECTS: THE 2008 REVISION, *Population Database*. Retrieved from $http : //esa.un.org/unpp/index.asp?panel = 2$

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

An Improved Information Rate of Perfect Secret Sharing Scheme Based on Dominating set of Vertices

N. M. G. Al-Saidi[1], N.A. Rajab[1], Mohamad Rushdan Md. Said[2] and K. A. Kadhim[1]

[1] *The Branch of Applied Mathematics Applied Sciences Department, University of Technology, Baghdad, Iraq*

[2] *Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, Darul Ehsan, 43400 Serdang, Malaysia*

emails: nadiamg08@gmail.com, nuha.abd10@gmail.com, rushdan@math.upm.edu.my, kadhum.technology@gmail.com

Abstract

Due to the fast development in data communication systems and computer networks in recent years, the necessity to protect the secret data is of great demands. Several methods have been arisen to protect the secret data; one of them is the secret sharing scheme. It is a method of distributing a secret K among a finite set of participants, in such a way that only predefined subset of participant is enabled to reconstruct a secret from their shares. A secret sharing scheme realizing uniform access structure described by a graph has received a considerable attention, where each vertex represents a participant and each edge represents a minimum authorized subset. In this paper, an independent dominating set of vertices in a graph G is introduced and applied as a novel idea to construct a secret sharing scheme such that the vertices of the graph represents the participants and the dominating set of vertices in G represents the minimal authorized set. This new scheme is based on principle of non-adjacent vertices, whereas, most of the previous works are based on the principle of adjacent vertices. We prove that the scheme is perfect, and the lower bound of the information rate for this new construction is improved as compared to some well-known previous constructions.

Key words: Secret sharing scheme, independent dominating set, information rate, uniform access structure, rank

1 Introduction

Secret sharing scheme is a method of distributing a secret K among a finite set of participants $P = \{p_1, \cdots, p_n\}$, such that, only predefined specific subset belonging to the access structure when they pooling their partial information together, can reconstruct the secret, whereas, any other subset not in the access structure can determine nothing about the secret. The first kind of secret sharing scheme called (t, n)-threshold scheme is introduced independently by Shamir [1] and Blakley [2] in 1979. A considerable attention was given to this subject after on. An efficient secret sharing scheme is the one that has high information rate (small share size) which is considered as a measure for the efficiency of such systems. Hence, an extensive consideration has been given to improve this value. It is defined as the ratio of the length of the secret to the average length of the share given to the participant. A special participant D, not in the set P, is called the dealer takes the responsibility to distribute the shares to each participant. The family of all the subset of participant, called an access structure Γ [3].

In the (t, n)-threshold scheme, all the t-subsets of a set of n-participant capable of reconstructing the secret are called the authorized sets and represented by Γ_0, whereas, those not belonging to Γ are called unauthorized sets. An authorized set B is called minimal if it satisfies the following relations; for $B' \subset B$, and $B' \in \Gamma$ implies $B' = B$. If $B \in \Gamma$ and $B \subset B'$ implies that $B' \in \Gamma$, then the access structure is called monotone. In this case, the collection of minimal authorized set could uniquely determine the access structure. Many approaches for construction of secret sharing scheme have been proposed. A special construction class is the threshold scheme. Ito, Saito, and Nishizeki [4] have generalized this concept and proposed a method for any monotone access structure, when the size of each share is no less than the size of the secret. This scheme is called perfect. It is called ideal, if the size of each share is the same as the size of the secret [5].

The relation between secret sharing scheme and an access structure have considerable interest by many researchers (see, for example [6, 7, 8]). Several approaches were developed to obtain an efficient implementation of secret sharing scheme for an arbitrary access structure, but there is no general method. An approach based on graph theory that investigates an access structure over a set of two participants are of special interest from the practical point of view, because such access structure can be represented by a graph, in which the vertices represent the participant, and the edges represent an access structure. This approach is called graph access structure and has been studied by many researchers (see, for example [4, 7, 8, 9]).

The general idea behind the B-decomposition of a graph access structure proposed by Stinson [8, 10], is to find all subgraphs of G, such that; at least B subgraphs of them have to contains each edge of E. This leads to an increase in the number of such subgraphs that increased exponentially with the number of vertices (participants), and following in an exponential time construction. In this paper, we propose an approach that investigates the minimum access structure in a manner differs from the previous construction that was based on adjacent principle of vertices. In our approach, the minimum authorized subsets represent the set of all minimum independent dominating set of vertices in the graph G.

The current work is organized as follows: In Section 2, some terminologies related to domination in graph G, and to secret sharing scheme is presented. The main algorithm for the decomposition and reconstruction of the proposed method is given in Section 3. The efficiency of the constructed secret sharing scheme is demonstrated and proved through the information rate in section 4. Finally, the paper is concluded in section 5.

2 Domination in graph and secret sharing scheme

Many theories and applications for the domination in graph theory as a natural model for many location problems have been arisen. As an example; the fire station problem, school bus routing, computer communication networks, social network theory, and coding theory. Some graph theoretical definitions, and some important terminologies related to secret sharing scheme are given to fill in some background for the reader. A more detailed review of the topics in this section, can be found in [11, 12, 13, 14, 15].

A graph G is an ordered pair $G = (V; E)$ where V is a set of elements which are called vertices, and the set $E = \{e = \{u, v\} \in E : u, v \in V\}$ is called edges. The order of G, $|G|$ equals $|V|$. A graph $H = (W; F)$ is a subgraph of another graph $G = (V; E)$ if $W \subseteq V$ and $F \subseteq E$. All graphs used in this work are simple, connected, and undirected.

Definition 2.1: The set of vertices adjesnt to $v \in V$ is called an open neighborhood $N(v)$ such that $N(v) = \{u \in V : \{u, v\} \in E\}$, whereas, the closed neighborhood is the set $N[v] = N(v) \cup \{v\}$.

Definition 2.2: The degree $d(v)$ of a vertex v is the size of $N(v)$, or equivalently, the number of edges incident to v.

Definition 2.3: A graph G is an r-regular if the minimum degree of a graph G equals its maximum degree and equals r. A 2-regular, 3-regular, and $(n - 1)$-regular graphs are called a cycle, cubic, and complete graphs respectively.

Definition 2.4: A graph G is decomposable into $H_1, H_2, \ldots, H_k$ if G has subgraphs $H_1, H_2, \cdots, H_k$, such that:

1. Each edge of G belongs to one of the H_i 's for some $i = 1, 2, \cdots, k$.
2. H_i and H_j have no edges in common, when $i \neq j$.

Definition 2.5: For a graph $G = (V; E)$, a set of vertices D is called a dominating set of vertices in G, if $N[D] = V$, or equivalently, every vertex in $V - D$ is adjacent to at least one vertex in D. A dominating set D is minimal if no proper subset of D is a dominating set.

Definition 2.6: An independent dominating set of vertices (ID) in a graph G, is the dominating set D that has no two vertices of D connected by an edge of G. The minimum cardinality of an independent dominating set ID is called a minimum independent dominating set MID where $|MID|$ is called minimum independent domination number. In this work, such type of dominating set is used in the construction for the secret sharing scheme.

The information rat that specifies the efficiency of the secret sharing schemes has introduced for the first time by Brickell [16]. He defined it as follows.

Definition 2.7: The information rate ρ of a secret sharing scheme is $\rho = log_2(|K|)/log_2(|S|)$, where K, and S, represents the secret and the share respectively. The term ideal is used for a perfect secret sharing scheme with information rate equals (1).

Definition 2.8: The maximum cardinality of a minimal qualified subset is called the rank (m) of an access structure Γ.

3 The proposed algorithm

A novel construction algorithm of a perfect secret sharing scheme with an access structure of rank m is proposed in this section. It is based on the same concepts given by H. Sun [17, 18]. This algorithm consists of three parts as follows.

1. The Initialization Phase
   (a) Given $G = (V, E)$, be an r-regular graph, with $V = \{v_1, v_2, \cdots, v_n\}$. Let $P = \{p_1, p_2, \cdots, p_n\}$ be the set of participant corresponding to the set V.
   (b) Construct Γ_0 by computing all (MID) in G, such that $\Gamma_0 = \{MID : V = \cup MID\}$. Hence, $|MID| = m$ the rank (or, the minimum qualified set).
   (c) All the computation is done over $GF(q)$, where q is a prime, and $q \geq 2n + 2$.
2. The Decomposition Phase
   (a) Decompose the graph G into n-subgraphs G_i, where $V(G_i) = \{V/N[v_i] : i = 1, 2, \cdots, n\}$.

(b) Decompose Γ_0 into n of Γ_i's, such that $\Gamma_i = \{MID \in \Gamma_0 \text{ where } MID \supseteq p_i\}$ and $\Gamma_0 = \bigcup_{i=1}^{n} \Gamma_i$.

(c) Define $\Gamma_i^* = \{X : X \cup p_i \in \Gamma_i\}$. The closure of Γ_i^* is a uniform access structure of rank $m-1$ or the set of all minimum independent set of vertices in the subgraph G_i.

(d) Let $K = \{k_1, k_2, \cdots, k_m\}$ be a secret, such that, k_i is taken randomly from $GF(q^{(m-1)!})$.

- A polynomial $f(x)$ of degree $(m-1)$ is selected by the dealer, where its coefficients are $K = \{k_1, k_2, \cdots, k_m\}$. Hence, $f(x) = (k_1 x^{m-1} + k_2 x^{m-2} + \cdots + k_m) \bmod q^{(m-1)!}$.
- Compute y_i, such that, $y_i = f(i)(modq), i = 1, 2, \cdots, n$. The secret is reconstructed by getting m or more y_i's.
- The dealer select n random numbers $r_1, r_2, \cdots, r_n$ over $GF(q^{(m-1)!})$.

(e) The value $r_i + y_i$ is distributed to each Γ_i^*. Hence the share of the participant p_i is as follows:

$$S_i =< r_i, S_i(\Gamma_1^*), \cdots, S_i(\Gamma_{i-1}^*), (\Gamma_{i+1}^*), \ldots, S_i(\Gamma_n^*) >.$$

The secret can be reconstructed when the participant of the authorizer pool their share together. The reconstruction phase could be expressed as follow:

3. The reconstruction phase

The secret can be reconstructed by getting m or more y_i. This can be done by combining their values together, for all $X \in \Gamma_0$ such that $X = (p_1, p_2, \cdots, p_m)$. We can obtain each y_i from $S_j(\Gamma_l^*)$, because Γ_l^* elements dominates the subgraph G_l. Hence we can obtain m of y_i that recovers $f(x)$ easily.

4 Improved information rate

Keeping the length of a share as small as possible is a desired purpose. This leads to an improved information rate that could result in an efficient and practical secret sharing scheme. The characteristics of the graph used in the construction, such as, order, regular, and graph dominating, etc., have affected the information rate values. Applying the proposed algorithm in Section 3, the relations between these characteristics are deduced. In this work, the experimental implementation is summarized in some important relations. These relations are observed using Table(1). It tabulates the relations between graph order, regularity, and graph domination. Based on these values an important generalized relation is abstracted in theorem (1). It describes the relation between the main characteristics of a graph with

Table 1: The relation between order, regular, and dominating set

$No.$	$Ver's.$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	$r=8$	——
1	4	MID=2	-	-	-	-	-	-	-
2	5	MID=2	-	-	-	-	-	-	-
3	6	MID=2	MID=2	MID=2	-	-	-	-	-
4	7	MID=3	-	MID=2	-	-	-	-	-
5	8	MID=3	MID=2	MID=2	MID=2	MID=2	-	-	-
6	9	MID=3	-	MID=2	-	MID=2	-	-	-
7	10	MID=4	MID=3	MID=2	MID=2	MID=2	MID=2	MID=2	-
8	11	MID=4	-	MID=3	-	MID=2	-	MID=2	-
9	12	MID=4	MID=3	MID=3	MID=2	MID=2	MID=2	MID=2	-
10	13	MID=5	-	MID=3	-	MID=2	-	MID=2	-
11	14	MID=5	MID=4	MID=3	MID=3	MID=2	MID=2	MID=2	-
12	15	MID=5	-	MID=3	-	MID=3	-	MID=2	-
13	16	MID=6	MID=4	MID=4	MID=3	MID=3	MID=2	MID=2	-
14	17	MID=6	-	MID=4	-	MID=3	-	MID=2	-
15	18	MID=6	MID=5	MID=4	MID=3	MID=3	MID=3	MID=2	-
16	19	MID=7	-	MID=4	-	MID=3	-	MID=3	-
17	20	MID=7	MID=5	MID=4	MID=4	MID=3	MID=3	MID=3	-
18	21	MID=7	-	MID=5	-	MID=3	-	MID=3	-
19	22	MID=8	MID=6	MID=5	MID=4	MID=4	MID=3	MID=3	-
20	23	MID=8	-	MID=5	-	MID=4	-	MID=3	-
21	24	MID=8	MID=6	MID=5	MID=4	MID=4	MID=3	MID=3	-
-	-	-	-	-	-	-	-	-	-

rank (m).

Observation 1.

Let $G=(V,E)$ be an r-regular simple graph with $|V|=n$, where $n\geq 4$ then $|MID|=m=\lceil\frac{n}{r+1}\rceil$.

Theorem 1.

Let $G=(V,E)$ be an r-regular simple graph with $|V|=n$, where $n\geq 4$ if $|MID|=m$, then,

$$m(r+1)\geq n\geq\begin{cases}m(r+1)-r & \text{if } r\,is\,even\\ m(r+1)-r+1 & \text{if } r\,is\,odd\end{cases}$$

Proof: Let G be an r-regular simple graph with $|V| = n$, $n \geq 4$ and $|MID| = m$, let $MID = \{v_1, v_2, \cdots, v_m\}$, such that $d(v_i) = r$, that represent the degree of the vertex v_i, and $N(v_i)$ represents the neighborhood of the vertices v_i, $i = 1, \cdots, m$.

To prove the upper bound, if $(N(v_i)\colon\ i = 1, \cdots, m)$ are disjoint sets, then $|V| = mr+m$. If $u \in \{N(v_i) \bigcap N(v_j\}, where\, i, j \in \{1, 2, \cdots, n\}$, that means, there is at least one vertex in common in two neighborhood. Therefore, $|V| < mr + m$.

To prove the lower bound, if r is an even number, and $N(v_i)$ are equal for all $i = 1, \cdots, m$, then $|V| = r + m$, this is true when $m = 2$. If $m > 2$ then there exists $N(v_i) \neq N(v_j)$ i.e. $(\exists u \notin \{N(v_i) \bigcap N(v_j)\}$, for any $i, j \in \{1, 2, \cdots, m\})$. Since the graph is r-regular, then any two different neighborhoods have at most r vertices in common. Therefore, $|V| > m(r + 1) - r$.
If r is an odd number, and at least two sets of neighborhoods common in a vertex v. Hence, the sets of neighborhood have at most $(r - 1)$ vertices in common, because there is no r-regular graph of odd order. Therefore, $|V| > rm + m - r + 1$. □

The general information rat are introduced using the notion of entropy, such that, the information rate of a secret sharing scheme that defined in [19], is as follows:

$$\rho = \frac{H(S)}{max_{i\in\{1,\cdots,n\}} H(I_i)} = \frac{log_2|S|}{max_{i\in\{1,\cdots,n\}} log_2(S_i)}. \tag{1}$$

Theorem 2.
Let $G = (V, E)$ be an r-regular simple graph with $|V| = n$, where $n \geq 4$, if $|MID| = m$, for $m \geq 4$, then the information rate of the secrete sharing graph of G is

$$\rho = \frac{(m-1)!\, log_2(q)}{max_t\{\Sigma_{(t:p_i\in G_t)}(|G_t| - N(p_i)) + 1\} log_2(q)} \tag{2}$$

Where G_t is the graph decomposition of G, $t = 1, 2, \cdots, n$.

Proof: To prove the information rate; since the share of participant p_i is $S_i =< r_i, S_i(\Gamma_1^*), \cdots, S_i(\Gamma_{i-1}^*), (\Gamma_{i+1}^*), \ldots, S_i(\Gamma_n^*) >$, and the length of $S_i(\Gamma_j^*)$ is equal to $log(p^{(|G_t|-N(p_i))})$, therefore, if $p_i \in G_t$, and $N(p_i)$ is the neighborhood of the vertex p_i, then, the length of share S_i is equal to $log({}_q(t : p_i \in G_t)^{\Sigma(|G_t|-N(p_i))+1})$. Since the length of the secret K is equal to $log(q^{(m-1)!})$, Hence, the information rate of the secret sharing scheme is ρ, such that;

$$\rho = \frac{log_2|K|}{log_2|S|} = \frac{(m-1)! log_2(q)}{max_t\{\Sigma_{(t:p_i\in G_t)}(|G_t| - N(p_i)) + 1\} log_2(q)} \tag{3}$$

The lower bound is calculated from the worst case when G_t is an r-regular subgraph. Since the number of vertices in G_t is $(n-r-1)$, then the lower bound is as follows,

$$\rho \geq \frac{(m-1)!}{((n-r)(n-r-1)+1)}, when, \ (m-1) < \frac{n}{(r_{Min}+1)} \leq m \tag{4}$$

The upper bound is as follows,

$$\rho \leq \frac{(m-1)!}{((n-r)(n-r-1)+1)}, when, \ (m-1) < \frac{n}{(r_{Max}+1)} \leq m \tag{5}$$

□

The average information rate with two bounds (upper, lower) is compared to the bounds given by H. Sun [17] in order to show that the constructed values have improved bounds. The comparison can be shown in Table (2).

Table 2: Comparison between our proposed method and Sun's [17] Method

No. of vertices n	Sun's method	Our method
$2r+2 \geq n \geq r+2$	$\rho \geq \frac{2}{(r+1)}$	$\rho \geq \frac{2}{r}$
$3r+1 \geq n \geq 2r+4$	$\rho \geq \frac{6}{((n-1)^2+2)}$	$\rho \geq \frac{6}{((n-r)(n-r-1)+2)}$
$m(r+1) \geq n \geq \begin{cases} m(r+1)-r & \text{if } r\, is\, even \\ m(r+1)-r+1 & \text{if } r\, is\, odd. \end{cases}$	$\rho \geq \frac{(n-m+1)}{\binom{n}{m}}$	$\rho \geq \frac{(m-1)!}{(n-r)(n-r-1)+1}$

Example: Let G be a graph with $n=8$, r=2, and m=3, then, based on the proposed algorithm in Section 3, the decomposition of the graph G is shown in Figure(1), where $\Gamma_0 = \{\{1,4,7\},\{2,5,8\},\{3,6,1\},\{2,4,7\},\{2,5,7\},\{3,5,8\},\{4,6,1\}\}$.

The information rate of the present proposed method is calculated using Table (2). This table presents the lengths of the shares. Therefore, by applying equation (2), the information rate of the proposed method equals ($\rho = 0.1875$), whereas, ($\rho = 0.1176$) using Sun's method.

Conclusions

A new decomposition construction for perfect secret sharing scheme with graph access structure is proposed in this paper. It is based on an independent dominating set of vertices in a graph G. In most of the previous proposed decomposition construction, the minimum

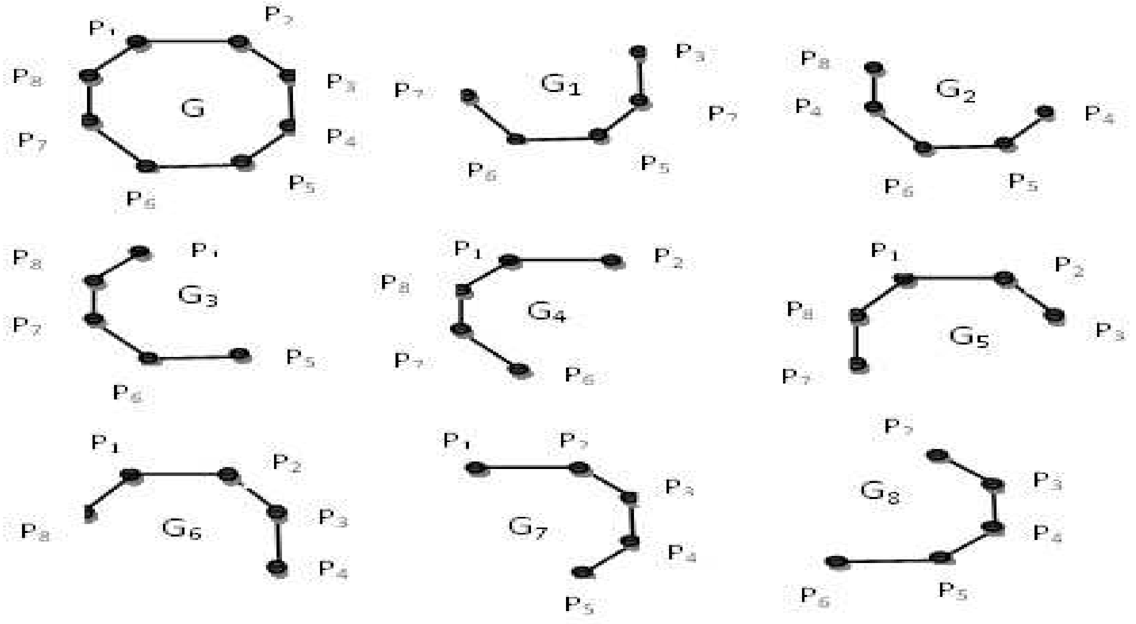


Figure 1: Decomposition based on dominating set of Vertices in Graph G

Table 3: The length of the shares for example 1

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
i=1	-	-	$log(q^4)$	$log(q^3)$	$log(q^3)$	$log(q^3)$	$log(q^4)$	-
i=2	-	-	-	$log(q^4)$	$log(q^3)$	$log(q^3)$	$log(q^3)$	$log(q^4)$
i=3	$log(q^4)$	-	-	-	$log(q^4)$	$log(q^3)$	$log(q^3)$	$log(q^3)$
i=4	$log(q^3)$	$log(q^4)$	-	-	-	$log(q^4)$	$log(q^3)$	$log(q^3)$
i=5	$log(q^3)$	$log(q^3)$	$log(q^4)$	-	-	-	$log(q^4)$	$log(q^3)$
i=6	$log(q^3)$	$log(q^3)$	$log(q^3)$	$log(q^4)$	-	-	-	$log(q^4)$
i=7	$log(q^4)$	$log(q^3)$	$log(q^3)$	$log(q^3)$	$log(q^4)$	-	-	-
i=8	-	$log(q^4)$	$log(q^3)$	$log(q^3)$	$log(q^3)$	$log(q^4)$	-	-

authorized subset is constructed by a set of edges in graph G, which is considered as small share in the construction of large schemes, whereas, in our scheme, the minimum authorized subset represents the minimum independent dominating set of vertices in the graph G. The information rate that used as a measure of the efficiency for secret sharing scheme is calculated. We conclude that the proposed scheme is the most efficient.

Acknowledgements

This work has been totally supported by University of Technology -applied sciences department-branch of applied mathematics -Baghdad-Iraq, and partially supported as a short term study by University Putra Malaysia-Institute for mathematical researches (INSPEM).

References

[1] A. Shamir, *How to share a secret*, Commun. of the ACM, **22**(1979) 612–613.

[2] G. R. Blakley, *Safeguarding cryptographic keys*, Proc. of the 1979 AFIPS National Computer Conference of AFIPS Conference proceedings, AFIPS Press **48** (1979) 313–317.

[3] H. Sun, and B. Chen, *On the decomposition constructions for perfect secret sharing schemes*, Proceeding of Information and Communication Security, ICICS'97, Beijing, China, (1997) 11–14.

[4] M. Ito, A. Saito and T. Nishizeki, *Secret sharing scheme realizing general access structure*,Proc. IEEE Globecom '87 (1987) 99–102.

[5] H. Sun, *Decomposition Construction for secret sharing schemes with graph access structures in polynomial time*, SIAM Journal on Discrete Mathematics,**24**(2) (2010) 617–638.

[6] A. Beimel, T. Tassa, and E. Weinreb, *Characterizing ideal weighted threshold secret sharing*, SIAM J. Discrete Math., **22** (2008) 360397.

[7] G. Di Crescenzo and C. Galdi, *Hypergraph decomposition and secret sharing*, Discrete Appl. Math., **157** (2009) 928946.

[8] C. Blundo, A. De Santis, D. R. Stinson, and U. Vaccaro,*Graph decomposition and secret sharing schemes* J. of Cryptology, **8**(1)(1995) 39-64.

[9] M. Liu, L. Xiao, and Z. Zhang, *Multiplicative linear secret sharing schemes based on connectivity of graphs*, IEEE Trans. Inform. Theory, **53** (2007) 39733978.

[10] E. F. Brickell and D. R. Stinson, *Some improved bounds on the information rate of perfect secret sharing schemes*, J. Cryptology, **5**(1992)153-166.

[11] P. A. Dreyer, *Applications and variations of domination in graph*, Ph. D. thesis, The State University of New Jersey, 2000.

[12] G. J. Chang, *Algorithmic aspects of domination in graphs*, Handbook of Combinatorial Optimization, (D.-Z. Du and P. M. Pardalos eds.), **3**(2011) 339-405.

[13] J. M. Tarr, *Domination in graphs*, Master thesis, University of South Florida, 2010.

[14] C. Blundo, A. De Santis, L. Gargano, and U. Vaccaro, *On the information rate of secret sharing schemes*, Theoretical Computer Science, **154**(2)(1996) 283-306.

[15] D. R. Stinson, *New general lower bounds on the information rate of secret sharing schemes*, Lecture Notes in Computer Science, **740**(1993) 170-184.

[16] E. F. Brickell, *Some ideal secret sharing schemes*, Journal of Combin. Math. and Combin. Comput., **6**(1989) 105–113.

[17] H. Sun, *Recursive constructions for perfect secret sharing schemes*, Computers and Mathematics with Applications, **37**(1999) 87–96.

[18] H. Sun, and S. Shieh, *Constructing perfect secret sharing schemes for general and uniform access structures*, Journal of information science and engineering, **15**(1999) 679–689.

[19] S. Iftene, *Secret sharing schemes with applications in security protocols*, Ph.D. thesis, Al. I. Cuza University of Ia'si, 2007.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Testing for a class of bivariate exponential distributions

V. Alba-Fernández[1] and M.D. Jiménez-Gamero[2]

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

[2] *Department of Statistics and O.R., University of Sevilla, Spain*

emails: mvalba@ujaen.es, dolores@us.es

Abstract

Bivariate and multivariate exponential distributions are widely applied in several areas such as reliability, queueing systems or hydrology. Among this kind of distributions, the Moran-Downton exponential distribution is an appealing choice. Nevertheless, for the inferences to be sound, it must be checked if it can be reasonably assumed that this model provides an adequate description of the available data. Here, a goodness-of-fit test for the Moran-Downton exponential distribution is proposed. The test statistic exploits the analytically convenient formula of the characteristic function of this distribution and compares it with the empirical characteristic function of the sample. Large sample properties of the proposed test are studied. The finite sample performance of the proposed test is numerically studied. Finally, the test is applied to a real data set describing the joint distributions of rainfall events for two rain gauge stations in Spain.

Key words: Empirical characteristic function, goodness-of-fit, Moran-Downton distribution, Bootstrap distribution estimator
MSC 2000: AMS 62F03, 62F40.

1 Introduction

Bivariate exponential distributions refer to bivariate distributions with both marginal distributions being exponential. There is no unique extension of the univariate exponential distribution to the bivariate case and several kinds of bivariate exponential distributions have been proposed (see for example [6, Ch. 10] and [5, Ch. 47]). Besides the applications in the area of life testing, reliability or queueing systems, this sort of distributions has proved useful for modeling hydrological data, which are essentially non-negative and asymmetric ([6, 5]).

Among the hydrological data analyses, the study of rainfall characteristics has been an area of great interest for many researchers studying rainfall patterns in many countries of the world. The use of univariate distributions has been a standard approach in such rainfall analysis. However, the application of bivariate distributions has not been pursued to any great extent, despite it being more relevant to these problems. The pair-wise dependence between rainfall characteristics has been described in the literature by using some families of bivariate distributions. Some of them are derived by means of the copula method and assuming exponential marginals. Applications of these bivariate distributions to study the joint distributions of different rainfall variables can be found in [10] for modeling rainfall intensity, depth and duration in the Amite River basin in Lousiana, United States; also in [2], where Gumbel's type I bivariate exponential distribution was used to model rainfall intensity and duration in the Po basin, Italy.

However, other bivariate exponential distributions which are not based on copulas have been also used. One of these distributions is the Moran-Downton bivariate exponential distribution, MD distribution for short. For instance, in [8] the authors claimed that this distribution is an appealing choice to be used for pairs of hydrological quantities such as streamflow at two points of a river, or rainfall at two locations, or in [3] where the authors applied this distribution to drought data from the state of Nebraska, United States. However, in the mentioned applications, the adequacy of the model is concluded from the fitted marginals and from the observed correlation. But this is clearly erroneous because there exist several bivariate exponential distributions with correlated components and exponential marginals, and some important quantities depending directly on the joint distribution may have rather different values for different bivariate exponential distributions having the same correlation coefficient. Thus, a formal procedure to check if one specified model is appropriate to fit a data set is necessary, beyond testing exponentiality of each component. In this sense, the aim of this paper is to develop a goodness-of-fit (gof) test for the MD distribution.

2 A gof test for the MD distribution

A bivariate random variable $X = (X_1, X_2)$ is said to have a MD distribution, $X \sim MD(\lambda_1, \lambda_2, \rho)$, if it has the following probability density function (pdf)

$$f(x;\lambda_1,\lambda_2,\rho) = \frac{\lambda_1\lambda_2}{1-\rho}\exp\left(-\frac{\lambda_1 x_1 + \lambda_2 x_2}{1-\rho}\right) I_0\left(\frac{2\sqrt{\lambda_1 x_1 \lambda_2 x_2 \rho}}{1-\rho}\right),$$

for $x_1, x_2 > 0$, where $\lambda_1, \lambda_2 > 0$ and $0 \le \rho < 1$ are the parameters of the distribution and I_0 is the modified Bessel function of the first kind of order zero,

$$I_0(z) = \frac{1}{\pi}\int_0^{\pi} e^{z\cos(x)}dx = \sum_{k\geq 0}\frac{\left(\frac{1}{4}z^2\right)^k}{(k!)^2}.$$

Equivalently, we write $X \sim MD(\theta)$, with $\theta = (\lambda_1, \lambda_2, \rho) \in \Theta = \{\theta = (\lambda_1, \lambda_2, \rho),\, \lambda_1, \lambda_2 > 0,\, 0 \leq \rho < 1, \}$. Clearly, X_1 and X_2 are independent if and only if $\rho = 0$. The marginal pdfs of X_1 and X_2 are exponential with scale parameters λ_1 and λ_2, respectively, and ρ is the coefficient of correlation between X_1 and X_2.

Since the pdf and the cumulative distribution function (cdf) of the MD distribution do not have an easily tractable expression (see [6, pp. 436]), the application of classical gof tests based on the cdf may become rather cumbersome. Here, we propose a gof test based on the characteristic function, which is given by

$$\phi_0(t;\theta) = \frac{\lambda_1\lambda_2}{(\lambda_1 - it_1)(\lambda_2 - it_2) + \rho t_1 t_2},$$

$i = \sqrt{-1}$. With this aim, we will also make use of the following property: if $(X_1, X_2) \sim MD(\lambda_1, \lambda_2, \rho)$, then $(Y_1, Y_2) \sim MD(1, 1, \rho)$, where $Y_k = \lambda_k X_k$, $k = 1, 2$.

Therefore, for testing

$$\begin{array}{ll} H_0: & X \sim MD(\theta), \;\text{ for some } \theta \in \Theta_0, \\ H_1: & X \nsim MD(\theta), \;\; \forall \theta \in \Theta_0, \end{array} \tag{1}$$

where $\Theta_0 = \{\theta = (\lambda_1, \lambda_2, \rho),\, \lambda_1, \lambda_2 > 0,\, 0 < \rho < 1, \} \subset \Theta$, a reasonable test function is the following one,

$$\Psi = \begin{cases} 1, & \text{if } T_{n,w}(\hat{\theta}) \geq t_{n,w,\alpha}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $t_{n,w,\alpha}$ is the $1-\alpha$ percentile of the null distribution of $T_{n,w}(\hat{\theta})$,

$$T_{n,w}(\hat{\theta}) = n \int \left| \frac{\phi_n(t)}{\phi(t; 1, 1, \hat{\rho})} - 1 \right|^2 w(t)dt, \tag{3}$$

$\phi_n(t)$ is the ecf associated with $Y_1, \ldots, Y_n$, $t = (t_1, t_2)$,

$$\phi_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(it_1 Y_{j1} + it_2 Y_{j2}),$$

$\hat{\theta} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\rho})$ be a consistent estimator of θ, $Y_j = (\hat{\lambda}_1 X_{j1}, \hat{\lambda}_2 X_{j2})$, $1 \leq j \leq n$, and $w(t) \geq 0$ is a weight function on $\mathbb{R}^2$ with finite integral, $\int w(t)dt < \infty$, and an unspecified integral denotes integration over the whole space $\mathbb{R}^2$.

Note that $T_{n,w}(\hat{\theta})$ is invariant with respect to changes of scale in each component of X. As a consequence, the null distribution of $T_{n,w}(\hat{\theta})$ does not depend on the parameters λ_1 and λ_2. Thus to derive the null distribution we can assume that the data come from a $MD(1, 1, \rho)$ distribution.

The presence of $w(t)$ in the expression of $T_{n,w}(\hat{\theta})$ has two main purposes: for adequate choices of w, it renders the integral (3) finite and it also gives a readily computable closed form of $T_{n,w}(\hat{\theta})$. A problem with the test function Ψ defined in (2) is the calculation of the critical point $t_{n,w,\alpha}$. The exact null distribution of the test statistic $T_{n,w}(\hat{\theta})$ is unknown. A classical way to overcome this problem consist in approximating the null distribution of the test statistic by means of its asymptotic null distribution. When H_0 is true, $T_{n,w}(\hat{\theta})$ converges in law to a linear combination of independent χ^2_1 variates, where the weights in this linear combination depend in a very complicated way on the unknown true value of the parameter θ, and thus they are unknown. Therefore, the asymptotic null distribution of $T_{n,w}(\hat{\theta})$ does not provide a useful approximation. Because of this reason, we propose to approximate the null distribution of $T_{n,w}(\hat{\theta})$ through a parametric bootstrap estimator (see for example Meintanis and Swanepoel [7], Jiménez-Gamero et al. [4], Alba-Fernández et al. [1]).

We have study some properties of the proposed test and discussed some practical considerations related to the $w(t)$ and $\hat{\theta}$. The properties studied are asymptotic, that is to say, they are valid for large samples. The finite sample performance of the proposed test is numerically analyzed by means of a simulation study.

3 Application

As mentioned before, the MD distribution has been used to model rainfall at two locations. Here, we applied the proposed gof test for modeling precipitation data (in mm) from two rain gauge stations in the Eastern of Spain. The two rain gauge stations analyzed were Valencia (V) and Reus (R). Monthly total rainfall data from January of 2000 to June of 2012 (n =140 observations) were collected from the website of the Spanish State Meteorological Agency (AEMET, www.aemet.es). The left panel of Figure 1 displays the scatterplot of the data.

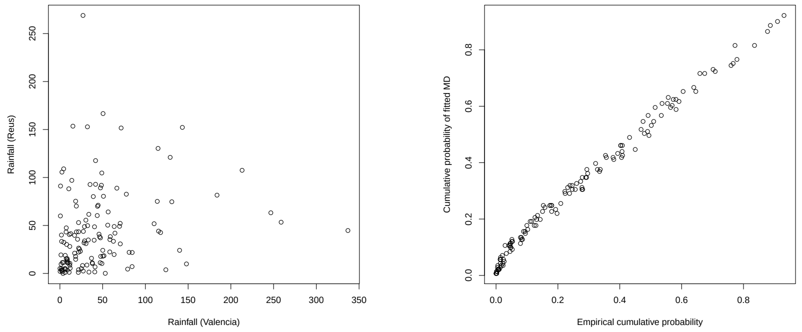


Figure 1: Scatterplot of Spanish data (left panel) and empirical versus theoretical cumulative probabilities (right panel).

Table 1: p-values for testing gof to MD.

p-values	$a = 0.5$	$a = 1$	$a = 2$
(V,R)	0.425	0.440	0.250

The estimations of the parameters are $\hat{\lambda}_1 = 0.0227$, $\hat{\lambda}_2 = 0.0234$, and $\hat{\rho} = 0.232$. The proposed test was applied to this data set, taking as weight function $w(t) = \exp(-a\|t\|^2)$, $a = 0.5, 1, 2$, $t \in \mathbb{R}^2$. The resultant p-values are shown in Table 1. From the obtained p-values, it can be concluded that the MD distribution provides a reasonable model for this data set. A graphical device that help us to see how well the MD distribution fits the data is that proposed by [9], that compares the joint cumulative relative frequencies to the theoretical cumulative probabilities. The right panel of Figure 1 displays it. Looking at this figure we see that the theoretical probabilities fit the empirical ones quite well, in fact, the associated correlation between them is 0.9942.

References

[1] M.V. Alba-Fernández, D. Barrera-Rosillo, M.J. Ibáñez-Pérez, M.D. Jiménez-Gamero, *A homogeneity test for bivariate random variables*, Comput. Stat. **24** (2009) 513–531.

[2] B. Bacchi, G. Becciu, N.T. Kottegoda, *Bivariate exponential model applied to intensities and durations of extreme rainfall*,Journal of Hydrology **155** (1994) 225–236.

[3] A.K. Gupta, S. Nadarajah, *Product moments of Downton's bivariate exponential distribution*, Water Resources Management **22** (2008) 671–679.

[4] M.D. Jiménez-Gamero, M.V. Alba-Fernández, J. Muñoz-García, Y. Chalco-Cano, *Goodness-of-fit tests based on empirical characteristic functions*, Comput. Stat. Data Anal. **53** (2009) 3957–3971.

[5] S. Kotz, N. Balakrishnan, N.L. Johnson, *Continuous Multivariate Distributions. Volume 1: Models and applications*, Wiley, 2000.

[6] N. Balakrishnan, C.D. Lai, *Continuous bivariate distributions*, Springer, 2009.

[7] S.G. Meintanis, J. Swanepoel, *Bootstrap goodness-of-fit tests with estimated parameters based on empirical transformations*, Statist. Probab. Lett. **77** (2007) 1004–1013.

[8] M. Nagao, M. Kadoya, *Two-variate exponential distribution and its numerical table for engineering application*, Bulletin of the Disaster Prevention Research Institute, Kyoto University **20** (1971) 183–215.

[9] S. Yue, *The bivariate lognormal distribution to model a multivariate flood episode*, Hydrological Processes **14** (2000) 2575–2588.

[10] L. Zhang, V.P. Singh, *Bivariate rainfall frequency distribution using Archimedean Copulas*,Journal of Hydrology **332** (2007) 93–109.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

A matrix algorithm for computing orbits in parallel and sequential dynamical systems

Juan A. Aledo[1], Silvia Martínez[1] and José C. Valverde[1]

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

emails: juanangel.aledo@uclm.es, silvia.msanahuja@uclm.es,
jose.valverde@uclm.es

Abstract

We provide a matrix method in order to compute orbits of parallel and sequential dynamical systems on maxterm and minterm Boolean functions. To be more precise, we study this problem when the dynamical system is defined over an undirected graph and also over a directed graph. We also extend the results to systems where the local Boolean functions are not dependent restrictions of a global one.

Key words: Discrete dynamical systems, parallel dynamical systems, sequential dynamical systems, computation of orbits, dependency graphs, Boolean functions.
MSC 2000: 37B99, 37E15, 37N99, 68R10, 94C10.

1 Summary

In the development of the theory of computer simulation, discrete dynamical systems have played an important role. Actually, in the last years, several computer processes have been mathematically modeled by dynamical systems (see [1, 2, 6, 11, 13, 14, 15]).

Certainly, in computer simulation, there are many entities and each entity has a state at a given time. Thus, the update of the states of the entities constitutes an evolution in time of the system, i.e., a discrete dynamical system (see [7, 8, 9]).

The update of the states is determined by relations of the entities, which can be represented by a dependency (directed or undirected) graph, and by local rules, which together constitute the (global) *evolution operator* of the dynamical system (see [12, 16]).

If the states of the entities are updated in a parallel manner, the system is called a *parallel dynamical system* (PDS) [1, 2, 6], while if they are updated in a sequential order, the system is named a *sequential dynamical system* (SDS) [6, 14, 15].

In a previous work [2], the authors proved that PDS over directed dependency graphs can present periodic orbits of any period when the evolution operators are general maxterms or minterms (see also [4]). This situation totally differs from the case of PDS over undirected dependency graphs, where only (eventually) fixed points or 2-periodic orbits can exist. Thus, depending on the structure of the (directed) dependency graph, the orbit structure of the system can turn out very involved due to the coexistence of periodic orbits of different periods.

According to [12] and [16], one of the main goals in the study of a dynamical system is to give a complete characterization of its orbit structure. In our particular case, as the state space of the system is finite, every orbit is periodic or eventually periodic. Nevertheless, taking into account the mentioned results in [2] and [3], determining a priori the different coexistent periods of its orbits seems to be impossible for an arbitrary system, since it depends both on the dependency graph and on the evolution operator.

It claims for a method to compute every orbit of the system, which is the purpose of this work. In fact, we develop algorithms in order to compute orbits of parallel and sequential dynamical systems defined over directed (and undirected) graphs when the evolution operator is a general minterm or maxterm and, likewise, when it is constituted by independent local Boolean functions, so providing a new tool for the study of orbits of these dynamical systems.

In this work we provide a matrix algorithm in order to compute orbits of parallel and sequential dynamical systems on maxterm and minterm Boolean functions. To be more precise, we study this problem when the dynamical system is defined over an undirected graph and also over a directed graph. We also extend the results to systems where the local Boolean functions are not dependent restrictions of a global one. All these results appear in our recent paper [5].

Acknowledgements

This work has been partially supported by the grants MTM2011-23221, PEII11-0132-7661 and PEII09-0184-7802.

References

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

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

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

[4] J.A. Aledo, S. Martinez and J.C. Valverde, *Parallel dynamical systems over special digraph classes*, Int. J. Comput. Math. DOI:10.1080/00207160.2012.742191

[5] J.A. Aledo, S. Martinez and J.C. Valverde, *Updating method for the computation of orbits in parallel and sequential dynamical systems*, Int. J. Comput. Math. DOI:10.1080/00207160.2013.767894

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

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

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

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

[10] E. A. Bender and S. G. Williamson, *A Short Course in Discrete Mathematics*, Dover Publications, Inc., Mineola, New York, 2005.

[11] O. Coln-Reyes, R. Laubenbacher and B. Pareigis, *Boolean monomial dynamical systems*, Ann. Combin. **8** (2004) 425–439.

[12] Y.A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer, New York, 2004.

[13] R. Laubenbacher and B. Pareigis, *Decomposition and simulation of sequential dynamical systems*, Adv. Appl. Math. **30** (2003) 655-678.

[14] H.S. Mortveit and C.M. Reidys, *Discrete, sequential dynamical systems*, Discrete Math. **226** (2002) 281–295.

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

[16] S. Wiggins, *Introduction to Applied Nonlinear Systems and Chaos*, Springer, New York, 1990.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Accelerating the Computation of Nonnegative Matrix Factorization in Multi-core and Many-core Architectures

P. Alonso[1], V.M. García[2], F.J. Martínez-Zaldívar[3], A. Salazar[3], L. Vergara[3] and A.M. Vidal[2]

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

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

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

emails: palonso@uniovi.es, vmgarcia@dsic.upv.es, fjmartin@dcom.upv.es, asalazar@dcom.upv.es, lvergara@dcom.upv.es, avidal@dsic.upv.es

Abstract

In this work, sequential and parallel implementations of the Lee & Seung algorithm for calculating the Nonnegative Matrix Factorization on parallel computers (with multi-core architecture and Graphics Processing Units) are presented. The algorithm has been reorganized to obtain a quality approach from two points of view: accuracy and execution time. Moreover, the performance of the proposed algorithms is analyzed.

Key words: NNMF, GPU, multi-core, many-core, parallel algorithm

1 Introduction

Let us consider the problem of approximating a nonnegative matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, i.e., all elements must be equal to or greater than zero, by means of a product of two nonnegative matrices $\mathbf{W} \in \mathbb{R}^{m\times k}$ and $\mathbf{H} \in \mathbb{R}^{k\times n}$, of low rank $k \leq min(m,n)$ with $\mathbf{A} \approx \mathbf{WH}$. Usually k is chosen to be smaller than m and n, so that $\mathbf{W}$ and $\mathbf{H}$ are smaller than the original matrix. The problem can be addressed as the computation of two matrices $\mathbf{W}_0$ and $\mathbf{H}_0$ such that

$$\|\mathbf{A} - \mathbf{W}_0\mathbf{H}_0\|_F = \min_{\mathbf{W}\geq\mathbf{0},\mathbf{H}\geq 0} \|\mathbf{A} - \mathbf{WH}\|_F \,, \tag{1}$$

where $\|\cdot\|_F$ denotes the well know Frobenius norm. (Although it is also possible to use another kind of distance instead of the Frobenius norm, throughout this paper we focus on the problem as outlined in Equation (1).)

The importance of this decomposition, known as Nonnegative Matriz Factorization (NNMF), lies in the fact that an approximation of the matrix **A** can be built from a positive basis **W** with a small number of columns, as compared with the size of the original matrix **A**. This allows that each column vector of **A** can be approximated as a linear combination of vectors of a base **W**, thus converting the NNMF in a very useful tool in many applications. In recent years, the NNMF has become an essential tool in fields such as document clustering, data mining, machine learning, data analysis, image analysis, audio source separation, bioinformatics, etc., [1], [3], [7], [18], [19].

It is necessary to have efficient algorithms that compute this decomposition in a short time, especially for those cases where the sizes of the matrices appearing in the problems are large enough to require a significant computational effort, or when there exist temporal restrictions in finding the solution of the problem. The use of new parallel architectures based on multi-core processors and/or Graphics Processing Units (GPUs) is a good alternative to reduce the computation time of the NNMF. Many algorithms have been developed for the calculation of this factorization, see for example [2], [5], [6], [8], [10], [11], [12], [16], [17], [20]. It can be remarked that presented in [12] by its simplicity and its ability to be effectively used in many applications.

Our principal goal is to develop a parallel algorithm that can be used in solving problems of audio source separation [9]. For this type of applications the use of parallel computers based on recent successful architectures (i.e. multi-cores or GPUs) is very adequate. Therefore we have focused on developing efficient parallel algorithms on such architectures. Previous experiences in parallelization of NNMF algorithms can be found in [4], where the authors developed an algorithm on GPUs. Also it is noteworthy the parallel approach in [14] for distributed memory clusters or the work described in [1] where the authors also focussed in an NNMF parallel algorithm for solving some audio signals problems.

In our case we have chosen to focus on the possibility of parallelization of an algorithm based on the Lee & Seung algorithm [12] on multi-core and GPUs architectures. This sequential algorithm provides very good performance in terms of simplicity and opportunities for parallelization. In this work we have tried to reorganize the algorithm in order to obtain a quality approach from the point of view of minimizing the error and to parallelize it in order to obtain a low cost per iteration.

The rest of the paper is organized as follows: Section 2 describes the sequential algorithm to be parallelized and analyzes its theoretical computational cost and some improvements to increase the accuracy of the algorithm. In Section 3, we discuss the parallelization strategy on multi-core and many-core architectures. Section 4 presents an experimental analysis of the performance of the parallel algorithm from two points of view: numerical accuracy and execution time improvements with regard to the sequential algorithm. Finally, we show some conclusions and future work in Section 5.

2 Sequential Lee & Seung Algorithm for computing NNMF

The simplest approach for calculating NNMF was provided by Lee & Seung in [12]. The idea is quite simple and only involves matrix-matrix product type operations and componentwise matrix operations. That is the reason why it is an algorithm easy to implement. In [12] it is also shown the convergence of this algorithm to a solution of Equation (1). The algorithm has the advantage of its simplicity but also the disadvantages of slow convergence towards the solution or the obtention of a high residue if a small number of iterations are carried out.

Algorithm 1 Lee & Seung Algorithm (LSA)

```
Inputs:
$\mathbf{A} \in \mathbb{R}^{m\times n}, A(i,j) \geq 0\ \ \forall i,j$
$k \leq \min(m,n)$
Outputs:
 $\mathbf{W} \in \mathbb{R}^{m\times k},\ \mathbf{H} \in \mathbb{R}^{k\times n},\ W(i,j), H(i,j) \geq 0\ \ \forall i,j$
 1: Choose random $\mathbf{W} \in \mathbb{R}^{m\times k}, \mathbf{H} \in \mathbb{R}^{k\times n},\ W(i,j), H(i,j) \geq 0$
 2: for $cont = 1, 2, ... \rightarrow$ Convergence do
 3:   $\mathbf{B} = \mathbf{W}^T\mathbf{W}\mathbf{H}; \mathbf{C} = \mathbf{W}^T\mathbf{A};$
 4:   $\mathbf{D} = \mathbf{W}\mathbf{H}\mathbf{H}^T; \mathbf{E} = \mathbf{A}\mathbf{H}^T;$
 5:   for $i = 1, 2, ..., k$ do
 6:     for $j = 1, 2, ..., n$ do
 7:       $H(i,j) = \frac{H(i,j)\cdot C(i,j)}{B(i,j)}$
 8:     end for
 9:   end for
10:   for $i = 1, 2, ..., m$ do
11:     for $j = 1, 2, ..., k$ do
12:       $W(i,j) = \frac{W(i,j)\cdot E(i,j)}{D(i,j)}$
13:     end for
14:   end for
15: end for
```

The main operations involved in the Lee & Seung Algorithm (LSA) are matrix multiplications (lines 3 and 4) and the elementwise matrix-matrix multiplications/divisions (lines 7 and 12, just inside loops 5 and 10 respectively). This results in a computational cost that can be expressed in flops per iteration as:

$$T_{\text{SEQ}} = (4mnk + 2k(m+n)(k+1))\,(\text{flops/iter}) \approx \left(4mnk + 2k^2(m+n)\right)(\text{flops/iter}).$$

Other good approaches for calculating the NNMF are the alternating least squares methods [5], [7], [10]. One of the advantages of LSA is that the matrix solutions **W** and **H** can be used as inputs of a more costly alternating least squares method in order to accelerate its convergence.

One possibility to accelerate LSA and improve the error bounds for a fixed number of iterations, is to update matrix **H** with the newly computed version of the matrix **W**. Now, the algorithm can be expressed as:

Algorithm 2 Modified Lee & Seung Algorithm (MLSA)

```
Inputs:
$\mathbf{A} \in \mathbb{R}^{m\times n}, A(i,j) \geq 0 \ \forall i,j$
$k \leq \min(m,n)$
Outputs:
$\mathbf{W} \in \mathbb{R}^{m\times k}, \ \mathbf{H} \in \mathbb{R}^{k\times n}, \ W(i,j), H(i,j) \geq 0 \ \forall i,j$
Choose random $\mathbf{W} \in \mathbb{R}^{m\times k}, \mathbf{H} \in \mathbb{R}^{k\times n}, \ W(i,j), H(i,j) \geq 0$
for $cont = 1,2,... \rightarrow$ Convergence do
    $\mathbf{B} = \mathbf{W}^T\mathbf{W}\mathbf{H}; \mathbf{C} = \mathbf{W}^T\mathbf{A};$
    for $i = 1,2,...,k$ do
        for $j = 1,2,...,n$ do
            $H(i,j) = \frac{H(i,j)\cdot C(i,j)}{B(i,j)}$
        end for
    end for
    $\mathbf{D} = \mathbf{W}\mathbf{H}\mathbf{H}^T; \mathbf{E} = \mathbf{A}\mathbf{H}^T;$
    for $i = 1,2,...,m$ do
        for $j = 1,2,...,k$ do
            $W(i,j) = \frac{W(i,j)\cdot E(i,j)}{D(i,j)}$
        end for
    end for
end for
```

With this strategy, the cost per iteration remains the same but the number of iterations required for convergence is improved.

3 Parallelizing the Lee & Seung Algorithm

Since the highest cost operations in LSA are basically matrix-matrix products, the simplest approach to parallelize it is to use numerical linear algebra libraries. This option allows, first of all, validation of algorithms and realization of a performance analysis in terms of accuracy, number of iterations required for convergence, comparisons with other standard algorithms, etc. On the other hand, it provides a set of reference implementations, which allows to analyze other specific implementations, explicitly organized by the programmer, with regard to runtime, speedup or other parameters. The generic parallel algorithm is shown in Algorithm 3.

Algorithm 3 Parallel Modified Lee & Seung Algorithm (PMLSA)

```
Inputs:
$\mathbf{A} \in \mathbb{R}^{m\times n}, A(i,j) \geq 0\ \forall i,j$
$k \leq \min(m,n)$
Outputs:
 $\mathbf{W} \in \mathbb{R}^{m\times k}$, $\mathbf{H} \in \mathbb{R}^{k\times n}$, $W(i,j), H(i,j) \geq 0\ \forall i,j$
Choose random $\mathbf{W} \in \mathbb{R}^{m\times k}, \mathbf{H} \in \mathbb{R}^{k\times n}$, $W(i,j), H(i,j) \geq 0$
for $cont = 1,2,... \rightarrow$ Convergence do
  $\mathbf{B} = \mathbf{W}^T\mathbf{W}\mathbf{H}; \mathbf{C} = \mathbf{W}^T\mathbf{A}$; (using parallel version of _gemm)
  for $i = 1,2,...,k$, (using implemented parallel version ---multithreaded/kernel---) do
    for $j = 1,2,...,n$ do
      $H(i,j) = \frac{H(i,j)\cdot C(i,j)}{B(i,j)}$
    end for
  end for
  $\mathbf{D} = \mathbf{W}\mathbf{H}\mathbf{H}^T; \mathbf{E} = \mathbf{A}\mathbf{H}^T$; (using parallel version of _gemm)
  for $i = 1,2,...,m$ (using implemented parallel version ---multithreaded/kernel---) do
    for $j = 1,2,...,k$ do
      $W(i,j) = \frac{W(i,j)\cdot E(i,j)}{D(i,j)}$
    end for
  end for
end for
```

In the case of the parallel approach for dense matrices on multi-core processors, a threaded version of BLAS3/LAPACK have been used to design the parallel implementation. This has allowed the efficient use of CPU resources with multiple cores, as well as automatic organization of blocks of parallel algorithms.

In the case of the implementation on GPUs for dense matrices, we have utilized the CUBLAS library [15]. As in the case of LAPACK, it allows obtaining good reference algorithms for the SIMD (Single Instruction, Multiple Data) organization of the algorithms implicit in this library.

In Algorithm 3, `_gemm` denotes `cublasDgemm/dgemm` for `CUBLAS/LAPACK` general matrix multiplications.

4 Experimental analysis

We analyze in this section the performance of the proposed algorithms from two points of view: accuracy and speedup. The machine where the tests were run has one Intel(R) Core i7-3820 CPU working at a frequency of 3.6 GHz with 10 240 kB of cache memory and 16 GB of main memory with hyperthreading capability, running a 64 bits Centos 6.4 Linux operating system with a 4.4.7 `gcc` version.

Besides, this machine incorporates a NVIDIA Kepler K20c GPU with 2496 cores and 4.7 GB of global memory. Its compute capability is 3.5 and the installed driver, SDK and toolkit version is 5.0.

Every result has been obtained with five tests (matrix $\mathbf{A}$ is uniformly distributed in the interval (0,1) and the initial values for $\mathbf{W}$ and $\mathbf{H}$ are the first k columns and rows of $\mathbf{A}$ respectively), discarding the best and the worst values and averaging the remaining intermediate ones. The double precision matrix multiplications have been computed using the Intel MKL Numerical Linear Algebra library release 11.0 threaded version using either one thread or the maximum (four —eight with hyperthreading—) number of threads. In the case of GPUs, matrix operations have been performed by using the CUDA CUBLAS library, version 5.0.

Table 1 summarizes some of the experimental results for algorithms based on the use of the aforementioned libraries using different versions. For several sizes of matrices ($m = n =$100, 500, 1000, 2000 and 4000, with $k = m/2$) it is shown: root mean squared residual r (Matlab R2012b [13] implementation of NNMF algorithm using the multiplicative update version with the default number of 100 maximum iterations and our proposed MLSA version with 100 iterations too), runtime in ms (GPU version and 1 thread and maximum number of thread versions) and speedup (GPUs version versus 1 thread and maximum number of threads versions).

It can be observed that the root mean squared residual r of the proposed MLSA method is always less than the Matlab residual and it decreases when the problem size increases for both methods. Besides, the speedup remains approximately constant with the dimension of the problem when it becames moderately large with values around 35 and 12 when either one thread or four threads respectively are used in the CPU.

$m = n$		100	500	1000	2000	4000
$r = \frac{\lvert\lvert A-WH \rvert\rvert_F}{\sqrt{m \cdot n}}$	(Matlab)	0.160	0.206	0.227	0.288	0.288
	(MLSA)	0.157	0.191	0.199	0.204	0.205
T_{GPU} (ms)		24	117	558	3516	26 231
T_{CPU} (ms)	(1 thread)	35	2321	16 201	120 679	935 508
	(max. threads)	26	1141	6922	45 172	324 523
Speedup	(1 thread)	1.5	19.8	29.0	34.3	35.7
	(max. threads)	1.1	9.8	12.0	12.8	12.4

Table 1: Performance

5 Conclusion

Sequential and parallel implementations of (M)LSA have been developed for calculating the NNMF on parallel computers with multicore and GPU architectures. Although there are currently more powerful algorithms, this algorithm is still competitive in terms of accuracy, ease of implementation and good performance on parallel computers. It is also possible to use matrices obtained by this algorithm as input matrices to other more powerful algorithms such as alternating least squares methods.

The performance obtained by the algorithms developed are promising considering that the acceleration in the calculation of the NNMF is not done at the expense of losing precision. In this work we focus on algorithms based on the use of libraries for dense matrices (such as LAPACK for multi-core architectures or CUBLAS for many-core architectures). The sparse case, a more detailed description of the performance of dynamic type algorithms on GPUs, or comparisons with other more powerful algorithms are lines of work that will be addressed shortly.

Acknowledgements

This work has been partially funded by PROMETEO/2009/013 project of Generalitat Valenciana, and projects TEC2009-13741 of the Ministerio Español de Ciencia e Innovación, and TEC2012-38142-C04 of the Ministerio Español de Economía y Competitividad.

References

[1] E. Battenberg, A.Freed, D. Wessel, *Advances In The Parallelization Of Music And Audio Applications.*, Ann Arbor, MI: MPublishing, University of Michigan Library 2010.

[2] M. Berry, M. Browne, A. Langville, V. Pauca, R. Plemmons, *Algorithms and applications for aproximate nonnegative matrix factorization*, Comput. Statist. Data Anal. **52** (2007) 155–173.

[3] A. Bjorck, *Numerical Methods for Least Squares Problems*, SIAM, North-Holland, Amsterdam, 1996.

[4] J.J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, F.J. Cañadas-Quesada, *Musical Instrument Sound Multi-Excitation Model for Non-Negative Spectrogram Factorization*, IEEE Journal of Selected Topics in Signal Processing **5(6**) (2011) 1144–1158.

[5] A. Cichocki, S. Cruces, S.I. Amari, *Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization*, Entropy **13**(2011) 134–170.

[6] A. Cichocki , A.H. Phan, *Fast local algorithms for large scale nonnegative matrix and tensor factorizations*, IEICE Trans. Fundamentals Electron. Commun. Comput. Sci. **E92- A**(2009), 708–721.

[7] A. Cichocki, R. Zdunek, S.I. Amari, *Hierarchical ALS algorithms for nonnegative matrix and 3d tensor factorization*, Lecture Notes in Comput. Sci. **4666** (2007) 169–176.

[8] N. Guan, D. Tao, Z. Luo, B. Yuan, *NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization*, IEES Transactions on Signal Processign **60(6)** (2012) 2882–2898.

[9] J. Kim, H. Park, *Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis*, Bioinform. **23** (2007) 1495–1502.

[10] J. Kim, H. Park, *Fast Nonnegative Matrix Factorization: An Active-Set-Like Method and Comparisons*, SIAM J. Sci. Comput. **33(6)**(2011) 3261–3281.

[11] D.D. Lee, H.S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature **401** (1999) 788–791.

[12] D.D. Lee, H.S. Seung, *Algorithms for non-negative matrix factorization*, MIT Press: Cambridge, MA **13** (2001) 556–562.

[13] MATLAB version 8.0 (R2012b), The MathWorks, Inc., Natick, Massachusetts, United States.

[14] E. Mejia-Roa, C. Garcia, J.I. Gomez, M. Prieto, C. Tenllado, A. Pascual-Montano, F. Tirado, *Parallelism on the Non-negative Matrix Factorization*, Applications, Tools and Techniques on the Road to Exascale Computing, vol. 22, Amsterdam, Netherlands, IOS Press BV, pp. 421–428, 2012.

[15] Nvidia, *CUBLAS Library Users Guide*, 2013.

[16] P. Paatero, U. Tapper *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics **5** (1994) 111–126.

[17] P. Paatero, *Least squares formulation of robust non-negative factor analysis*, Chemometrics Intell. Lab. Syst. **37**(1997) 23–35.

[18] F.J. Rodriguez-Serrano, J.J. Carabias-Orti, P. Vera-Candeas, T. Virtanen, N. Ruiz-Reyes, *Multiple Instrument Mixtures Source Separation Evaluation Using Instrument-Dependent NMF Models*, LNCS **7191** (2012) 380–387.

[19] J. Wnag, W. Zhong, J. Zhang, *NNMF-Based Factorization Techniques for High-Accuracy Privacy Protection on Non-negative-valued Datasets*, Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on Computing & Processing. pp. 513–517, 2006.

[20] W. Xu, X. Liu, Y. Gong , *Document Clustering Based On Non-negative Matrix Factorization*, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR 2003, pp. 267–273, 2003.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Almost Strictly Sign Regular matrices

Pedro Alonso[1], **Juan Manuel Peña**[2] **and María Luisa 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

In this work we announce a characterization of almost strictly sign regular matrices using Neville elimination. Neville elimination is an elimination procedure that consists of making zeros in a column of a matrix by adding to each row an adequate multiple of the previous one.

Key words: Almost Strictly Sign Regular matrices, Neville elimination, characterization
MSC 2000: AMS codes (optional)

1 Introduction

Totally positive (TP) and sign regular (SR) matrices arise naturally in many areas of mathematics, statistics, economics, etc (see [3], [4], [10] and [11]). On the other hand, in some recent papers it has been shown that an elimination procedure, called Neville elimination, is very convenient when working with totally positive and sign regular matrices. Many properties of these matrices and its subclasses have been proved using this elimination procedure (see [6] and [7]).

Roughly speaking, Neville elimination consists of making zeros in a column of a matrix by adding to each row an adequate multiple of the previous one (see [5] for more details), instead of using just a row with a fixed pivot as in Gauss elimination. This process is an alternative to Gaussian elimination which has been proved to be very useful with totally positive matrices, sign regular matrices or other related types of matrices, without increasing the error bounds (see [2]). Furthermore, we show that using this procedure it is possible to obtain algorithms with high relative accuracy (HRA) for the computation of eigenvalues and inverses of Pascal matrices (total nonnegative matrices) (see [1]). Other related results on HRA can be seen in [9].

In the last years different authors have studied the matrices strictly totally positive (STP) matrices, almost strictly totally positive (ASTP) matrices (see [7]), strictly sign regular (SSR) matrices or almost strictly sign regular (ASSR) matrices (see [8]).

In this work we announce some results about ASSR matrices. The main result characterizes ASSR matrices using Neville elimination.

2 Basic notations and auxiliary results

For $k, n \in \mathbb{N}$, with $1 \leq k \leq n$, $Q_{k,n}$ denotes the set of all increasing sequences of k natural numbers not greater than n. For $\alpha = (\alpha_1, \ldots, \alpha_k)$, $\beta = (\beta_1, \ldots, \beta_k) \in Q_{k,n}$ and A an $n \times n$ real matrix, we denote $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows $\alpha_1, \ldots, \alpha_k$ and columns $\beta_1, \ldots, \beta_k$ of A. If $\alpha = \beta$, we denote by $A[\alpha] := A[\alpha|\alpha]$ the corresponding principal minor. $Q^0_{k,n}$ denotes the set of increasing sequences of k consecutive natural numbers not greater than n.

First we define the matrices type-I and type-II staircase.

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

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

Definiton 2 *A matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ is called type-II staircase if it satisfies simultaneously the following conditions*

- $a_{1,n} \neq 0,\ a_{2,n-1} \neq 0,\ \ldots, a_{n,1} \neq 0$;
- $a_{ij} = 0,\ j > n - i + 1 \Rightarrow a_{kl} = 0,\ \forall i \leq k,\ j \leq l$;
- $a_{ij} = 0,\ j < n - i + 1 \Rightarrow a_{kl} = 0,\ \forall k \leq i,\ l \leq j$.

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 denote

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

for $k = 1, \dots, l$:

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

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

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

Analogously we denote

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

for $k = 1, \dots, r$:

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

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

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

Finally, we denote by I, J, $\widehat{I}$ and $\widehat{J}$ the following sets of indices, thereby defining the zero pattern in the matrix A:

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

Definiton 3 *Given a matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ type-I (type-II) staircase, we say that a submatrix $A[\alpha|\beta]$, with $\alpha, \beta \in Q_{m,n}$ is nontrivial if all its main diagonal (secondary diagonal) elements are non-zero.*

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

Next, we define the SR, SSR and ASSR matrices, and finally we present the characterization performed in [8] for ASSR matrices (see Theorem 10, pp 4184).

Definiton 4 *Given a vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) \in \mathbb{R}^n$, we say that ε is a signature sequence, or simply, is a signature, if $|\varepsilon_i| = |\pm 1| = 1$, $\forall i \in \mathbb{N}$, $i \leq n$.*

Definiton 5 *A real matrix A $n \times n$ is said to be SR, with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, if all its minors satisfy that*

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

Definiton 6 *A real matrix A $n \times n$ is said to be SSR, with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, if all its minors satisfy that*

$$\varepsilon_m \det A[\alpha|\beta] > 0, \quad \alpha, \beta \in Q_{m,n}, \qquad m \leq n. \tag{8}$$

Definiton 7 *A real matrix A $n \times n$ is said to be ASSR, with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, if all its nontrivial minors $\det A[\alpha|\beta]$ satisfy that*

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

Theorem 1 *Let A a real matrix $n \times n$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ 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*

$$\varepsilon_m \det A[\alpha|\beta] > 0. \tag{10}$$

3 Main results

Let $A = (a_{ij})_{1 \leq i,j \leq n}$ be a $n \times n$ matrix and $h = 1, \dots n-1$, we denote by A_h the matrix defined as

$$A_h := (a^h_{ij})_{1 \leq i,j \leq n-h+1}, \quad a^h_{ij} := a_{i+h-1,j+h-1}. \tag{11}$$

Analogously, the transpose of A_h, i.e. A^T_h, is denoted as

$$A^T_h := (a^{T,h}_{ij})_{1 \leq i,j \leq n-h+1}, \quad a^{T,h}_{ij} := a_{j+h-1,i+h-1}. \tag{12}$$

In the following two results, we announce necessary conditions for nonsingular ASSR type-I and type-II staircase matrices, respectively.

Theorem 2 *Let $B = (b_{ij})_{1 \leq i,j \leq n}$ be a nonsingular type-I staircase matrix, with zero pattern defined by I, J, $\widehat{I}$, $\widehat{J}$. If B is ASSR with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, then the Neville elimination of B can be performed without row exchanges and the pivots p_{ij} satisfy, for any $1 \leq j \leq i \leq n$,*

$$p_{ij} = 0 \Leftrightarrow b_{ij} = 0 \tag{13}$$

$$\varepsilon_{j-j_t} \varepsilon_{j-j_t+1} p_{ij} > 0 \Leftrightarrow b_{ij} \neq 0 \tag{14}$$

where $\varepsilon_0 = 1$,

$$j_t = \max \{ j_s \ / \ 0 \leq s \leq k-1, \ j - j_s \leq i - i_s \} \tag{15}$$

and k is the only index satisfying that $j_{k-1} \leq j < j_k$.

Theorem 3 *Let $B = (b_{ij})_{1 \leq i,j \leq n}$ be a nonsingular type-II staircase matrix, with zero pattern defined by I, J, $\widehat{I}$, $\widehat{J}$. If B is ASSR with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$, then the Neville elimination of B^T can be performed without row exchanges and the pivots q_{ij} satisfy, for any $1 \leq i \leq j \leq n$,*

$$q_{ij} = 0 \Leftrightarrow b_{ij} = 0 \tag{16}$$

$$\varepsilon_{i-\widehat{i}_t}\varepsilon_{i-\widehat{i}_t+1}q_{ij} > 0 \Leftrightarrow b_{ij} \neq 0 \tag{17}$$

where $\varepsilon_0 = 1$,

$$\widehat{i}_t = \max\left\{\widehat{i}_s \;/\; 0 \le s \le k-1,\; i-\widehat{i}_s \le j-\widehat{j}_s\right\} \tag{18}$$

and k *is the only index satisfying that* $\widehat{i}_{k-1} \le i < \widehat{i}_k$.

Finally, in the following two results we characterize all ASSR matrices through Neville elimination.

Theorem 4 *A nonsingular matrix* $A = (a_{ij})_{1\le i,j\le n}$ *is ASSR with signature* $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, *with* $\varepsilon_2 = 1$ *if and only if for every* $h = 1, \ldots, n-1$ *the following properties hold simultaneously:*

(i) *A is type-I staircase;*

(ii) *the Neville elimination of the matrices* A_h *and* A_h^T *can be performed without row exchanges;*

(iii) *the pivots* p_{ij}^h *of the Neville elimination of* A_h *satisfy conditions corresponding to (13), (14), and the pivots* q_{ij}^h *of the Neville elimination* A_h^T *satisfy (16) and (17);*

(iv) *for the positions* (i^h, j^h) *of matrix* A_h*:*

- *if* $i^h \ge j^h$ *and* $i^h - j^h = i_t^h - j_t^h$ *then* $\varepsilon_{j^h-j_t^h}\varepsilon_{j^h-j_t^h+1} = \varepsilon_{j^h-1}\varepsilon_{j^h}$,
- *if* $i^h < j^h$ *and* $i^h - j^h = \widehat{i}_t^h - \widehat{j}_t^h$ *then* $\varepsilon_{i^h-\widehat{i}_t^h}\varepsilon_{i^h-\widehat{i}_t^h+1} = \varepsilon_{i^h-1}\varepsilon_{i^h}$,

where indices $i_t^h, j_t^h, \widehat{i}_t^h, \widehat{j}_t^h$ *are given by conditions corresponding to (15) and (18).*

Theorem 5 *Let* P_n *be the* $n \times n$ *backward identity matrix. A nonsingular matrix* $A = (a_{ij})_{1\le i,j\le n}$ *is ASSR with signature* $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, *with* $\varepsilon_2 = -1$ *if and only if for every* $h = 1, \ldots, n-1$ *the following properties holds simultaneously:*

(i) $B = P_nA$ *is type-I staircase;*

(ii) *the Neville elimination of the matrices* $B_h = P_nA_h$ *and* $B_h^T = P_nA_h^T$ *can be performed without row exchanges;*

(iii) *the pivots* p_{ij}^h *of the Neville elimination of* B_h *satisfy conditions corresponding to (13), (14), and the pivots* q_{ij}^h *of the Neville elimination of* B_h^T *satisfy (16) and (17);*

(iv) *for the positions* (i^h, j^h) *of matrix* P_nA_h*:*

- *if* $i^h \ge j^h$ *and* $i^h - j^h = i_t^h - j_t^h$ *then* $\varepsilon_{j^h-j_t^h}\varepsilon_{j^h-j_t^h+1} = \varepsilon_{j^h-1}\varepsilon_{j^h}$,
- *if* $i^h < j^h$ *and* $i^h - j^h = \widehat{i}_t^h - \widehat{j}_t^h$ *then* $\varepsilon_{i^h-\widehat{i}_t^h}\varepsilon_{i^h-\widehat{i}_t^h+1} = \varepsilon_{i^h-1}\varepsilon_{i^h}$,

where indices $i_t^h, j_t^h, \widehat{i}_t^h, \widehat{j}_t^h$ *are given by conditions corresponding to (15) and (18).*

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. Delgado, R. Gallego, J. M. Peña, *Conditioning and accurate computations with Pascal matrices*, J. Comput. Appl. Math, in press.

[2] P. Alonso, J. Delgado, R. Gallego, J. M. Peña, *Growth Factors of Pivoting Strategies Associated to Neville Elimination*, J. Comput. Appl. Math, **235 (7)** (2011) 1755–1762 .

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

[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, *On the characterization of almost strictly totally positive matrices*, Adv. Comput. Math. **3(3)** (1995) 239–250.

[7] M. Gasca, J. M. Peña, *Characterizations and decompositions of almost strictly positive matrices*, SIAM J. Matrix Anal. Appl. **28(1)** (2006) 1–8.

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

[9] P. Koev, *Accurate eigenvalues and SVDs of totally nonnegative matrices*, SIAM J. Matrix Anal. Appl. **27** (2005) 1–23.

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

[11] A. Pinkus, *Totally Positive Matrices*, Cambridge University Press, Cambridge, UK, 2010.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

An ECC based key agreement protocol for mobile access control

J.A. Alvarez-Bermejo[1], **M.A. Lodroman**[2] **and J.A. Lopez-Ramos**[2]

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

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

emails: jaberme@ual.es, antonela_lodroman@yahoo.com, jlopez@ual.es

Abstract

We introduce a key agreement protocol based on the Diffie-Hellman key exchange protocol for elliptic curves that do not need the existence of a PKI infrastructure. The protocol is suitable light devices with low computational and storage capabilities where mobile devices can directly authorize other mobile devices to exchange keys to access a service of system in a secure an authenitcated manner.

Key words: Elliptic Curves, Diffie-Hellman key exchange
MSC 2000: AMS codes (optional)

1 Introduction

Mobile authentication in an insecure channel is nowadays an important issue due to the huge growth of applications for mobile devices in our society. Although authentication can be implemented using traditional public key cryptography ([7], [2]), due to the low computational resources of most mobile devices, cryptosystems using modular exponentation are rarely implemented, so this traditional cryptography is being substituted by the so called Elliptic Curve Cryptosystems (ECC) introduced in [3] and [4], mainly due the new standards for public key cryptography recommended in [5].

ECC shows significant advantadges concerning key sizes and fast computations, so it shows to be a great alternative due to its higher efficiency and low requirements for key length. Thus an ECC based authentication protocol shows to be more suitable for mobile devices.

However ECC also needs a public key infrastructure to maintain certificates and when the number of user increases, the store requirements grow and also the number of such certificates. Moreover, when implementing an authenticated key agreement protocol, the corresponding public key of each party should be authenticated. Therefore, since deploying a PKI system is not easy, some alternatives has been considered as identity based key agreement protocols as those given in [1], [6] or [8].

Most of them have been proven to be unsecure against some kind of attack, but the main problem that we have to face off is efficiency. Our aim in this paper is to introduce a new authenticated key agreement protocol based on the Diffie-Hellman problem for elliptic curves where users can play different roles and trusting is based on the roles based by these users.

2 The protocol

The protocol is divided into four different phases: initialization, key generation, key registration, authentication and key revocation and where participants may play four different roles: administrator, super user, visitor user and guard.

There exist two basic rules in the protocol: on one hand we have the key owners, that will be *administrator*, *super users* and the *visitor users*; on the other hand, the key verifiers, that will be the *guards*. Among the keys we will have three different grades: those belonging to the administrator and super user will be called *master keys* and the others will be *visitor keys*. Thus the roles played for every participant will determine the attributes of each one, i.e., access to a determined system and/or capability to generate other keys. Concerning this second issue, those holding a master key will be allowed to generate other master key or a visitor key and those holding a visitor key will not be allowed to generate any key.

2.1 Initialization

Each guard system $\mathcal{G}$ will match at least one administrator $\mathcal{A}$. The initialization phase will be developed under a secure channel and will consist on the following steps:

1. $\mathcal{A}$ chooses an elliptic curve $\mathcal{E}$, $P \in \mathcal{E}$ and an integer $r_{\mathcal{A}}$ and computes $Q_{\mathcal{A}} = r_{\mathcal{A}}P$.
2. $\mathcal{A}$ sends a message to $\mathcal{G}$ consisting in $Q_{\mathcal{A}}||P||\mathcal{E}||ID_{\mathcal{A}}$.
3. $\mathcal{G}$ chooses $r_{\mathcal{G}}$, computes $Q_{\mathcal{G}} = r_{\mathcal{G}}P$ and stores $K_{\mathcal{AG}} = r_{\mathcal{G}}Q_{\mathcal{A}}$
4. $\mathcal{G}$ sends back a message to $\mathcal{A}$ consisting in $ID_{\mathcal{G}}||Q_{\mathcal{G}}$.
5. $\mathcal{A}$ stores $K_{\mathcal{AG}} = r_{\mathcal{A}}Q_{\mathcal{G}}$

Thus the common key among $\mathcal{A}$ and $\mathcal{G}$ will be $K_{\mathcal{AG}}$.

ISBN: 978-84-616-2723-3

2.2 Key generation

When a user $\mathcal{U}$ wishes access to any system or service whose guard is $\mathcal{G}$ then he demmands either a master key or a visitor key to an administrator $\mathcal{A}$ or a super user $\mathcal{S}$. Let us assume that the request is sent to $\mathcal{A}$, who is sharing $K_{\mathcal{AG}}$. This phase is given by the following steps:

1. $\mathcal{U}$ chooses an elliptic curve $\mathcal{E}$, $P \in \mathcal{E}$ and an integer $r_{\mathcal{U}}$ and computes $Q_{\mathcal{U}} = r_{\mathcal{U}}P$.
2. $\mathcal{U}$ sends a message to $\mathcal{G}$ consisting in $KeyGen||Q_{\mathcal{U}}||P||\mathcal{E}||ID_{\mathcal{G}}||ID_{\mathcal{U}}$.
3. $\mathcal{A}$ chooses $r_{\mathcal{A}}$ and computes $Q_{\mathcal{A}} = r_{\mathcal{A}}P$, $K_{\mathcal{AS}} = r_{\mathcal{A}}Q_{\mathcal{U}}$, $C = E_{K_{\mathcal{AG}}}(Q_{\mathcal{U}}, ID_G, ID_A, P, \mathcal{E}, VT_{\mathcal{U}})$ and $H_{\mathcal{A}} = E_{K_{\mathcal{AS}}}(C)$, where $VT_{\mathcal{U}}$ will determine a valid time to use the demmanded key. $\mathcal{A}$ may also store $K_{\mathcal{AS}}$ for future secure communications with $\mathcal{U}$
4. $\mathcal{A}$ sends a message to $\mathcal{U}$ consisting in $C||H_{\mathcal{A}}||Q_{\mathcal{A}}$.
5. $\mathcal{U}$ computes $K_{\mathcal{AS}} = r_{\mathcal{U}}Q_{\mathcal{A}}$ and $D = E_{K_{\mathcal{AS}}}(C)$, verfies that $D = H_{\mathcal{A}}$ and stores C.

2.3 Key Registration

In this phase the user $\mathcal{U}$ will demmand access to the system or service whose guard is $\mathcal{G}$ via the information C that $\mathcal{U}$ received from $\mathcal{A}$ and at the end, both $\mathcal{G}$ and $\mathcal{U}$ will share a common key to either stablish secure communciations or $\mathcal{U}$ (depending on his role), can generate new keys for other users to give access to the system given by $\mathcal{G}$. The steps forming this phase are:

1. $\mathcal{U}$ chooses $Q \in \mathcal{E}$ and an integer $s_{\mathcal{U}}$ and computes $R_{\mathcal{U}} = s_{\mathcal{U}}Q$.
2. $\mathcal{U}$ sends a message to $\mathcal{G}$ consisting in $KeyReg||C||ID_{\mathcal{A}}||ID_{\mathcal{U}}||Q||R_{\mathcal{U}}||\mathcal{E}$.
3. $\mathcal{G}$ gets the information given in C using $K_{\mathcal{AG}}$, verifies its identity $ID_{\mathcal{G}}$, that $\mathcal{A}$ is authenticated with him, validity of VT_U and $\mathcal{E}$. He also gets $Q_{\mathcal{U}}$ and P.
4. $\mathcal{G}$ chooses two integers $u'_{\mathcal{G}}$ and $r'_{\mathcal{G}}$ and computes the following:
   - $R_{\mathcal{G}} = u'_{\mathcal{G}}Q$.
   - $S_{\mathcal{G}} = u'_{\mathcal{G}}R_{\mathcal{U}}$.
   - $K = h(S_{\mathcal{G}})$ for h a hash function.
   - $R'_{\mathcal{G}} = r'_{\mathcal{G}}P$.
   - $K_{\mathcal{UG}} = r'_{\mathcal{G}}W_{\mathcal{U}}$.

- $J = E_K(K_{\mathcal{U}\mathcal{G}})$.

5. $\mathcal{G}$ sends a message to $\mathcal{U}$ consisting in $R_{\mathcal{G}}||K||R'_{\mathcal{G}}||J$.

6. $\mathcal{U}$ computes $D = s_{\mathcal{U}} R_{\mathcal{G}}$ and $h(D)$

7. $\mathcal{U}$ verifies that $h(D) = K$.

8. $\mathcal{U}$ computes $F = s_{\mathcal{U}} R'_{\mathcal{G}}$ and $E_K(F)$, verifying that $E_K(F) = J$.

9. $\mathcal{U}$ stores $F = K_{\mathcal{U}\mathcal{G}}$.

Now $K_{\mathcal{U}\mathcal{G}}$ will be use either to authenticate U when trying to access the system whose guard is $\mathcal{G}$ as we will show in the next or to generate new master or visitor keys as introduced before.

2.4 Authentication

Authentication is the phase that grants access to any system or service supervised by $\mathcal{G}$. The guard $\mathcal{G}$ will authenticate any key owner $\mathcal{U}$ via their shared key using a typical challenge method.

1. $\mathcal{U}$ sends a message to $\mathcal{G}$ consisting in $Auth||ID_{\mathcal{U}}$.

2. $\mathcal{G}$ checks validity $T_{\mathcal{U}}$, i.e., time of access to the service for $\mathcal{U}$, chooses an integer r randomly and sends back the message $ID_{\mathcal{G}}||r$.

3. $\mathcal{U}$ computes and sends the message $n = E_{K_{\mathcal{A}\mathcal{G}}}(r)$.

4. $\mathcal{G}$ grants the access in case $n = E_{K_{\mathcal{A}\mathcal{G}}}(r)$.

2.5 Key Revocation

Key Revocation could be necesary as in the case of compromosing the master key of any user. Thus every key that was created from this user could also be compromised and so, all them should be revoked. The idea is then simple: we just have to represent every key in a directed graph where parent vertexes correspond to key generators and child vertexes to key requestor. Then if a key should be revoked as well as every key generated from the corresponding user to the first one, we will revoke this key and all the keys of users in its child vertexes.

3 Implementation aspects

As exposed in the previous sections, the usage of elliptic curves has the main advantage of being suitable for computation in embedded devices and sensor networks (mainly because the length of the keys is not as big in bytes- as it is in other cryptographic methods).

Implementing solutions where the higher range of devices is achieved, is a must in current developments; this, added to the need of focusing in low power consumption devices and mobile devices as they are the mainstream devices in market, such developments should be considered in an efficient and multiplatform language. Java, is a multiplatform programming language with strong components devoted to Internet connectivity, wich is a desirable issue- but it is also a definition for an abstract model for a computing platform, this has two direct consequences, the former is that the code written once (unless using special directives or libraries to use accelerators such as GPUs- or using assembler such as Oolong- to accelerate complex operations that lacks performance during the JIT stage) executes everywhere (including ARM embedded platforms, see [12]). And the latter, is that no matter what the device the code is ran on, the specification is optimized for it, and the code is run with warranties of being efficient in a high percentage. An example of such optimized virtual machines for embedded devices that enable ECC in them is the Squawk VM, which is a result of efforts to write a J2ME CLDC [10] compliant Java Virtual Machine, in Java, which provides OS level mechanisms for embedded devices, portability and debugging of the VM. It is primarily written in Java, and is compliant with the Connected Limited Device Configuration (CLDC) 1.1 Java Micro Edition (Java ME) configuration [11]. The fact that it runs without the need for an operating system is commonly known as running on the bare metal [9]. Nevertheless, when facing the arithmetic underneath the exposed protocols (namely, scalar and points products, inversions, signatures,) some optimizations may be necessary in order to allow embedded processors face the computation required, as in [12] where authors modify existing algorithms to use less storage space for precomputed values and to accelerate the execution in an ARM processor. As known, the performance of the elliptic curve cryptosystem (ECC) deeply depends on the computation of scalar multiplication. Thus, how to speed up the computation of the elliptic curve scalar multiplication is a significant issue. In [12], Lim and Lee, proposed a more flexible precomputation method used in wireless networks environments for speeding up the computation of exponentiation. This method can be also used for speeding up the scalar multiplication of elliptic curves. Authors, in [12], used the Java language to design the user interface of the ECDSA-like system, and their implementation experienced significative improvements. Therefore, an efficient method to secure information through ECC is feasible of being implemented in a very easy and reusable manner via the Java programming language and virtual machine, and being ported to a wide range of computing platforms ranging from supercomputers to mobile devices (ARM), sensor networks, and so on. Added to this, it is worth to cite the resources (libraries) available to developers to manage concurrence and security (message digest,) in an almost transparent way.

Acknowledgements

First author is supported by TIN2008-01117 grant of Ministerio de Ciencia e Innovación. Third authord is supported by Junta de Andalucía grant FQM 0211.

References

[1] Chen, L., Kudla, C., Identity based authenticated key agreement protocols from pairings, *Computer Security Foundations Workshop, 2003, Proceedings 16th IEEE,* 219–233, 2003.

[2] ElGamal, T., A public key cryptosystem and a signature scheme based on discrete logarithms, *Information Theory, IEEE Transactions* 31(4), 469–472, 1985.

[3] Koblitz, N., Elliptic curve cryptosystems, *Mathematics of Computation,* 48(177), 203–209, 1987.

[4] Miller, V., Use of elliptic curves in cryptography, *CRYPTO: Proceedings of Crypto,* 417–426, 1985.

[5] NSA suite b cryptography, http://www.nsa.gov/ia/programs/suiteb_cryptography/

[6] Okamoto, E., Proposal for identity-based key distribution systems, *Electronic Letters,* 22(24), 1283–1284, 1986.

[7] Rivest, R., Sharmir, A., Adleman, L., A method for obtaining digital signatures and public-key cryptography, *Communications of the ACM* 21(2), 120–128, 1978.

[8] Shim, K., Efficient id-based authenticated key agreement protocol based on weil pairing, *Electronics Letters,* 39(8), 653–654, 2003.

[9] Simon, D., Cifuentes, C., Cleal, D., Daniels, J., White, D., Java$^{\mathrm{TM}}$ on the bare metal of wireless sensor devices: the Squawk virtual machine, *Proceedings of the 2nd International Conference on Virtual Execution Environments 2006, Ottawa, Ontario, Canada,* 2006.

[10] Sun Microsystems, Inc., Java ME: Connected Limited Device Configuration (CLDC); JSR 30, JSR 139 (online), available: http://java.sun.com/products/cldc/, 2009 (accessed 19 February 2009).

[11] Sun Microsystems, Inc., Java ME Technology APIs and Docs (online), available: http://java.sun.com/javame/reference/apis.jsp, 2009 (accessed 19 February 2009).

[12] Tsaur, W.-J., Chou, C.-H., Efficient algorithms for speeding up the computations of elliptic curve cryptosystems, *Applied Mathematics and Computation,* 168(2), 1045–1064, 2005.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

The Conservative Method of Calculating the Boltzmann Collision Integral for Simple Gases, Gas Mixtures and Gases with Rotational Degrees of Freedom

Yu. A. Anikin[1,2], O. I. Dodulad[1,2], Yu. Yu. Kloss[1,2] and F. G. Tcheremissine[1,3]

[1] *, Moscow Institute of Physics and Technology*

[2] *, National Research Centre Kurchatov Institute*

[3] *, Dorodnicyn Computing Centre of RAS*

emails: , `dodulad@list.ru`, ,

Abstract

The conservative method of calculating the Boltzmann collision integral for simple gases, gas mixtures and gases with rotational degrees of freedom of molecules is presented. The method is applied to two fundamental rarefied gas problems: the heat transfer problem, and the problem of the shock wave structure.

Key words: Boltzmann equation, conservative method, gas mixture, gas with rotational degrees of freedom, shock wave, heat transfer

1 Introduction

Simulation of gas flows in microdevices and vacuum equipment, calculation of flows over aircraft in the upper atmosphere, analysis of processes in the shock front region and solution of many other problems, where the characteristic flow size is comparable with the mean free path of gas molecules, require application of kinetic theory methods and solution of the Boltzmann equation.

Numerical solution of this equation is a complicated task due to its high dimensionality and difficulties with calculating the multidimensional nonlinear collision integral. Today, the following methods of numerical solution or replacement of the Botzmann equation are known:

- statistical simulation methods [1]. They are characterized by significant statistical noise making it difficult to obtain highly accurate solutions.
- numerical analysis method based on expansion of the distribution function in polynomials and calculation of the collision integral from the approximation obtained. This method proved to be useful for one-dimensional shock wave structure problems [2], [3] and heat transfer problems [4], but it is very time-consuming and not conservative. The method has not been used for 2D or 3D flows.
- discrete-velocity methods that consider only those molecular collisions which initial and final velocities belong to the selected velocity grid. These methods allow obtaining results not affected by statistical noise as is the case of statistical simulation methods. The solution convergence to the exact solution of the Boltzmann equation at higher discretization of the velocity space has been proved [5], however, the convergence is slow, which calls for significant computing resources to obtain somewhat reliable results.
- spectral methods. The basic approaches are described in [7] and [6]. An attempt to resolve the spectral methods conservatism issue is presented in [8]. Earlier works consider only spatially uniform problems, but in recent years the spectral method was also applied to one-dimensional in the physical space problems.
- projection method [9]. This is a discrete ordinate method which key feature is the conservative projection method of computing the collision integral, which provides for strict fulfillment of conservation laws, non-negative solution, and zero value of the collision integral of the Maxwellian distribution function. The method has been used successfully for calculation of both supersonic flows [10] and slow flows only slightly deviating from the thermodynamic equilibrium state [11].

The present work is devoted specifically to the conservative projection method of calculating the Boltzmann collision integral. The method is described in the following section. It is presented as applied to a single-component monoatomic gas and generalized for a gas mixture. A modification of the conservative projection method, the multipoint projection method, is presented for a mixture of gases differing greatly in molecular masses. For diatomic molecules, there is described a method of solving the generalized Boltzmann equation that takes into account the exchange of translational and rotational energy of molecules. The method uses pre-calculated differential cross-sections of collisions of molecules with rotational degees of freedom, which makes the calculations more realistic compared to those presented in [12], where a two-stage model of inelastic collision and a probabilistic model of transitions between rotational levels [13] are considered.

Particular attention in the work is given to the use of realistic molecule interaction potentials, which allowed making a comparison between numerical results and experimental data. This is discussed in Section 3 that considers two fundamental rarefied gas problems: the heat transfer problem, and the problem of finding the shock wave structure.

2 Conservative projection method of calculating the collision integral

2.1 The method for a simple gas

The approximation of the collision integral is built for a simple gas as follows:

$$I(\xi) = \int \big(f(\boldsymbol{\xi}_1')f(\boldsymbol{\xi}') - f(\boldsymbol{\xi}_1)f(\boldsymbol{\xi}) \big)\, g\sigma(\theta, g)\, d^2\boldsymbol{n}\, d^3\boldsymbol{\xi}_1.$$

Here $\boldsymbol{\xi}$, $\boldsymbol{\xi}_1$ are pre-collision velocities of the molecules, $g = |\boldsymbol{g}|$, $\boldsymbol{g} = \xi_1 - \xi$ is the relative velocity vector, $\boldsymbol{\xi}' = 0.5g - 0.5g\boldsymbol{n}$, $\boldsymbol{\xi}_1' = 0.5g + 0.5g\boldsymbol{n}$ are post-collision velocities of the molecules, $\boldsymbol{n}$ is the unit vector of the direction of scattering, $\sigma(\theta, g)$ is the scattering cross-section that in the general case depends on the normal angle $\cos\theta = (\boldsymbol{n}, \boldsymbol{g})/g$ and the absolute value of relative velocity g. For the hard-sphere model the cross-section is a constant $\sigma(\theta, g) = 4\pi d^2$, where d is the hard sphere diameter.

When the Boltzmann equation is solved numerically in the velocity space, there is selected domain Ω of volume V where a grid of equally spaced velocity nodes $\{\xi_\gamma\}$, $\gamma = 1, \ldots, N$ is built. On the generated grid, the distribution function and collision integral are represented in the basis of δ-functions in the following form:

$$f(\xi) = \frac{V}{N} \sum_{\gamma=1}^{N} f_\gamma \delta(\xi - \xi_\gamma), \quad I(\xi) = \frac{V}{N} \sum_{\gamma=1}^{N} I_\gamma \delta(\xi - \xi_\gamma).$$

The symmetry property allows rearranging the collision integral into the form

$$I_\gamma = \int \big(\delta(\xi - \xi_\gamma) + \delta(\xi_1 - \xi_\gamma) - \delta(\xi' - \xi_\gamma) - \delta(\xi_1' - \xi_\gamma) \big) \times \\ \times \big(f(\boldsymbol{\xi}_1')f(\boldsymbol{\xi}') - f(\boldsymbol{\xi}_1)f(\boldsymbol{\xi}) \big)\, g\sigma(\theta, g)\, d^2\boldsymbol{n}\, d^3\boldsymbol{\xi}_1.$$

For calculation of the symmetrized collision integral, there is used a uniform 8-dimensional cubature Korobov grid [14] of nodes in space $\boldsymbol{\xi}_{\alpha_\nu}$, $\boldsymbol{\xi}_{\beta_\nu}$, $\boldsymbol{n}_\nu$, $\nu = 1, \ldots, N_\nu$ which nodes $\boldsymbol{\xi}_{\alpha_\nu}$, $\boldsymbol{\xi}_{\beta_\nu}$ coincide with the nodes of the velocity grid.

Since post-collision velocities are not in the grid nodes, the last two δ-functions in the projector are replaced with a combination of projectors into two pairs of nodes (ξ_{λ_ν}, $\xi_{\lambda_\nu+s_\nu}$) and (ξ_{μ_ν}, $\xi_{\mu_\nu-s_\nu}$) nearest to $\boldsymbol{\xi}_{\alpha_\nu}$ and $\boldsymbol{\xi}_{\beta_\nu}$, respectively, where s_ν is the 3D vector of displacement to the neighboring grid node, which components can take values of 0, 1, -1

$$\begin{aligned} \delta(\boldsymbol{\xi}' - \boldsymbol{\xi}_\gamma) &= (1 - r_\nu)\delta(\boldsymbol{\xi}_{\lambda_\nu} - \boldsymbol{\xi}_\gamma) + r_\nu \delta(\boldsymbol{\xi}_{\lambda_\nu+s_\nu} - \boldsymbol{\xi}_\gamma), \\ \delta(\boldsymbol{\xi}_1' - \boldsymbol{\xi}_\gamma) &= (1 - r_\nu)\delta(\boldsymbol{\xi}_{\mu_\nu} - \boldsymbol{\xi}_\gamma) + r_\nu \delta(\boldsymbol{\xi}_{\mu_\nu-s_\nu} - \boldsymbol{\xi}_\gamma). \end{aligned} \tag{1}$$

Such a replacement means that contributions to the collision integral in non-nodal points $\boldsymbol{\xi}'$ and $\boldsymbol{\xi}'_1$ are distributed among pairs of the nearest grid nodes. The coefficient r_ν is found in view of the energy conservation law $\xi^2_{\alpha_\nu}+\xi^2_{\beta_\nu}=(1-r_\nu)(\xi^2_{\lambda_\nu}+\xi^2_{\mu_\nu})+r_\nu(\xi^2_{\lambda_\nu+s_\nu}+\xi^2_{\mu_\nu-s_\nu})$.

The law of conservation of momentum is automatically fulfilled for the uniform grid with the symmetric choice of projection nodes as described above. The collision integral calculation formula takes the form

$$I_\gamma=\sum_{\nu=1}^{N_\nu}B_\nu(-(\delta_{\gamma,\alpha_\nu}+\delta_{\gamma,\beta_\nu})+(1-r_\nu)(\delta_{\gamma,\lambda_\nu}+\delta_{\gamma,\mu_\nu})+r_\nu(\delta_{\gamma,\lambda_\nu+s_\nu}+\delta_{\gamma,\mu_\nu-s_\nu}))\times \\ \times(f(\boldsymbol{\xi}'_1)f(\boldsymbol{\xi}')-f_{\alpha_\nu}f_{\beta_\nu})|\boldsymbol{\xi}_{\beta_\nu}-\boldsymbol{\xi}_{\alpha_\nu}|, \tag{2}$$

where $\delta_{\alpha,\beta}=\begin{cases}1,\ \alpha=\beta,\\ 0,\ \alpha\neq\beta\end{cases}$ is the Kronecker symbol, and $f_{\alpha_\nu}=f(\boldsymbol{\xi}_{\alpha_\nu})$, $f_{\beta_\nu}=f(\boldsymbol{\xi}_{\beta_\nu})$ are values of the grid distribution function.

Various interpolation types can be used to calculate $f(\boldsymbol{\xi}'_1)f(\boldsymbol{\xi}')$, but the preferred one is polynomial

$$f(\boldsymbol{\xi}'_1)f(\boldsymbol{\xi}')=(f_{\lambda_\nu}f_{\mu_\nu})^{1-r_\nu}(f_{\lambda_\nu+s_\nu}f_{\mu_\nu-s_\nu})^{r_\nu}, \tag{3}$$

that provides for zero value of the collison integral of the Maxwellian distribution function and fulfillment of the Boltzmanns H-theorem for the discretized equation.

At the time step τ the Boltzmann equation is solved using the symmetric scheme of splitting into the advection operator $T_{\Delta t/2}$ and relaxation collision operator $C_{\Delta t}$:

$$f_{\Delta t}=T_{\Delta t/2}(C_{\Delta t}(T_{\Delta t/2}(f_0))) \tag{4}$$

The relaxation equation is integrated using a scheme that has correct asymptotics at $t\to\infty$. Let us write the equation in the integral form, having introduced intermediate time values $t_\nu=t_0+\tau\nu/N_1$ and intermediate values of $f^{\gamma,\nu}$, where $f^{\gamma,\nu}|_{\nu=0}$, …, $f^{\gamma,\nu}|_{\nu=N_1}$ are distribution functions at time t_0, …, $t_0+\tau$, respectively

$$f^{\gamma,N_1}=f^{\gamma,0}+\int_{t_0}^{t_0+\tau}I^\gamma dt=f^{\gamma,0}+\sum_{\nu=1}^{N_1}\int_{t_0+\tau(\nu-1)/N_1}^{t_0+\tau\nu/N_1}I^\gamma dt, \tag{5}$$

which is rewritten as the continuous computation scheme

$$\begin{array}{ll} f^{\alpha_\nu,\nu}=f^{\alpha_\nu,\nu-1}-\tau B_\nu\Delta_\nu g_\nu, & f^{\beta_\nu,\nu}=f^{\beta_\nu,\nu-1}-\tau B_\nu\Delta_\nu g_\nu,\\ f^{\lambda_\nu,\nu}=f^{\lambda_\nu,\nu-1}+(1-r_\nu)\tau B_\nu\Delta_\nu g_\nu, & f^{\mu_\nu,\nu}=f^{\mu_\nu,\nu-1}+(1-r_\nu)\tau B_\nu\Delta_\nu g_\nu,\\ f^{\mu_\nu-s_\nu,\nu}=f^{\mu_\nu-s_\nu,\nu-1}+r_\nu\tau B_\nu\Delta_\nu g_\nu, & f^{\lambda_\nu+s_\nu,\nu}=f^{\lambda_\nu+s_\nu,\nu-1}+r_\nu\tau B_\nu\Delta_\nu g_\nu, \end{array} \tag{6}$$

where $\Delta_\nu=(f_{\lambda_\nu}f_{\mu_\nu})^{1-r_\nu}(f_{\lambda_\nu+s_\nu}f_{\mu_\nu-s_\nu})^{r_\nu}-f_{\alpha_\nu}f_{\beta_\nu}$, $B_\nu=\pi\sigma(\theta_\nu,g_\nu)VN/N_1$.

In problems where the interaction potential differs from the hard-sphere potential, collision cross-sections $\sigma(\theta,g)$ are calculated in advance by the trajectory method and stored in the form of grid dependence $\sigma(\theta_i,g_j)$. When the Boltzmann equation is being solved, the cross-section for specified θ,g is restored by interpolation of the grid dependence values $\sigma(\theta_i,g_j)$.

2.2 The projection method for a gas mixture

For a gas mixture, the Boltzmann equation takes the form:

$$\frac{\partial f_i}{\partial t} + \frac{\boldsymbol{p}}{m_i}\frac{\partial f_i}{\partial \boldsymbol{x}} = \sum_j \int \left(f_j(\boldsymbol{p}_1')f_i(\boldsymbol{p}') - f_j(\boldsymbol{p}_1)f_i(\boldsymbol{p})\right) g_{ij}\sigma(\theta, g_{ij})\, d^2\boldsymbol{n}\, d^3\boldsymbol{p}_1.$$

where f_i is the function of distribution of molecules of type i, $i = 1 \dots N$ by momenta.

The projection method described above can also be used for calculation of the collision integral for a gas mixture. The only requirement that shall be met for the sake of conservatism is the change from the velocity distribution function to the momentum distribution function, and the grid step in the momentum space $\Delta\boldsymbol{p}_i = \Delta\boldsymbol{p}$, $i = 1 \dots N$ for each component in the gas mixture.

Then, as in the case of the simple gas, each momentum after collisions is projected on two nodes in the grid, which makes the method conservative in terms of energy. Conservativeness in terms of momentum is achieved by symmetric choice of nodes in the grid.

The method works well at medium ratios of molecular masses of the mixture components (see [16], [15]), however, the equality of steps in momentum grids results in the increased number of nodes in the grids for heavy components of the mixture. The number of nodes increases according to relation $N_i \sim (m_i/m_{\min})^{3/2}$. For example, the calculation for a helium-argon mixture with the mass ratio $m^{\mathrm{He}}/m^{\mathrm{Ar}} = 1/10$ results in a 30-times increase of the grid, which increases dramatically the intensity of the calculations.

To remove this shortcoming of the method, let us consider momentum grids with uniform step $\Delta\boldsymbol{p}_i$ that however differs for each component, and perform projection of the contributions independently for each component of the mixture.

After presentation of the distribution function and collision integral in the basis of δ-functions, reduction of the collision integral to the symmetrical form, and application of the Korobov cubature grid $\boldsymbol{p}_{\alpha_\nu}, \boldsymbol{p}_{\beta_\nu}, \boldsymbol{n}_\nu$, the expression for calculation of the collision integral acquires the form

$$\begin{aligned} I_i(\boldsymbol{p}_\gamma) = \sum_j^N \sum_{\nu=1}^{N_\nu} B_\nu \left(\delta(p' - p_\gamma) + \delta(p_1' - p_\gamma) + \delta(p_{\alpha_\nu} - p_\gamma) - \delta(p_{\beta_\nu} - p_\gamma)\right) \times \\ \times \left(f_j(\boldsymbol{p}_1')f_i(\boldsymbol{p}') - f_i(\boldsymbol{p}_{\alpha_\nu})f_j(\boldsymbol{p}_{\beta_\nu})\right) \left|\frac{\boldsymbol{p}_{\alpha_\nu}}{m_i} - \frac{\boldsymbol{p}_{\beta_\nu}}{m_j}\right| \end{aligned} \tag{7}$$

where $\delta(p_{\alpha_\nu} - p_\gamma) = \delta_{\alpha_\nu,\gamma}$, $\delta(p_{\beta_\nu} - p_\gamma) = \delta_{\beta_\nu,\gamma}$, Kronecker symbols, and $f_i(\boldsymbol{p}_{\alpha_\nu}) = f_{i,\alpha_\nu}$, $f_j(\boldsymbol{p}_{\beta_\nu}) = f_{j,\beta_\nu}$ values of the grid distribution fucntion. As in the case of simple gas, the following expressions shall be determined by means of the projection procedure: $\delta(p' - p_\gamma)$, $\delta(p_1' - p_\gamma)$, $f_i(\boldsymbol{p}')$, $f_j(\boldsymbol{p}_1')$.

To provide for conservation of mass, three components of momentum and energy, the minimum number of projection nodes equals to five. The projection stencil used $\Lambda_5 =$

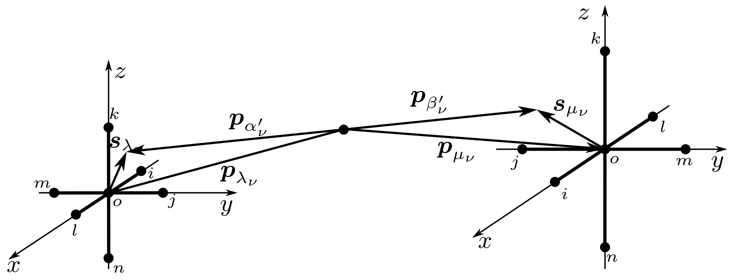


Figure 1: Projection stencil.

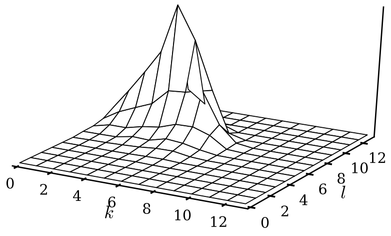


Figure 2: Differential cross-section as a function of l and k.

$\{0, i, j, k, 5\}$ is shown in Fig. 1, where $\boldsymbol{s}_{\lambda_\nu} = (\boldsymbol{p}_{\lambda_\nu} - \boldsymbol{p}')/\Delta p$, $\boldsymbol{s}_{\mu_\nu} = (\boldsymbol{p}_{\mu_\nu} - \boldsymbol{p}'_1)/\Delta p$, and $\boldsymbol{p}_{\lambda_\nu}$, $\boldsymbol{p}_{\mu_\nu}$ are momenta on the grid nearest to $\boldsymbol{p}'$, $\boldsymbol{p}'_1$.

It should be noted that all five projection nodes shall not be chosen from the vertices of the cube to which momentum $\boldsymbol{p}'$ belongs, for otherwise an inconsistent system of algebraic equations for projection coefficients will be obtained. The projection coefficients that define presentation of -functions in the form

$$\delta(p' - p_\gamma) = \sum_{a\in\Lambda} r_{\lambda,a}\delta_{\lambda_\nu+s_\nu,\gamma} \tag{8}$$

where s_a is the vector of displacement of the a-th stencil node, are $r_{\lambda,0} = 1 - \frac{1}{3}p - \frac{2}{3}q$, $r_{\lambda,5} = -\frac{1}{6}(p - q$, $r_{\lambda,i} = |s_{\lambda x}| + r_{\lambda,5}$, $r_{\lambda,j} = |s_{\lambda y}| + r_{\lambda,5}$, $r_{\lambda,k} = |s_{\lambda z}| + r_{\lambda,5}$, $p = |s_{\lambda x}| + |s_{\lambda y}| + |s_{\lambda z}|$, $q = s_{\lambda x}^2 + s_{\lambda y}^2 + s_{\lambda z}^2$.

Through direct substitution of contribution coefficients in expressions below, we can see that the scheme presented herein possesses the conservation property:

$$\sum_{a\in\Lambda} r_{\lambda,a} = 1, \ \sum_{a\in\Lambda} r_{\lambda,a}\boldsymbol{p}_{\lambda+s_a} = \boldsymbol{p}', \ \sum_{a\in\Lambda} r_{\lambda,a}p_{\lambda+s_a}^2 = p'^2.$$

The stencils can also be built on a larger number of neighboring nodes, which will additionally conserve the tensor of the flux of momentum and the energy flux vector at projection [17].

Values of distribution functions $f_i(\boldsymbol{p}')$ and $f_j(\boldsymbol{p}'_1)$ are found by interpolation of the nodal distribution function. The interpolation used is similar to the polynomial interpolation (2) and posseses the same property of zero collision integral of the Maxwellian function

$$f_i(\boldsymbol{p}') = \prod_{a\in\Lambda} (f(\boldsymbol{p}_{\lambda+s_a}))^{r_{\lambda,a}}. \tag{9}$$

Eventually, the formula for calculating the collision integral takes the form (7), where undefined expressions are calculated by formulas (8) and (9).

2.3 The projection method for a gas with rotational degrees of freedom

Behavior of a diatomic gas which molecules have rotational degrees of freedom is described at the kinetic level by the generalized Boltzmann equation [12]:

$$\frac{\partial f_i}{\partial t} + \boldsymbol{\xi}\frac{\partial f_i}{\partial \boldsymbol{x}} = \sum_{j,k,l}\int \Big(w_{ij}^{kl} f_l(\boldsymbol{\xi}_1') f_k(\boldsymbol{\xi}') - f_j(\boldsymbol{\xi}_1) f_i(\boldsymbol{\xi})\Big)\, g_{ij}\sigma(\theta, g_{ij})\, d^2\boldsymbol{n}\, d^3\boldsymbol{\xi}_1.$$

The equation is written for the distribution function $f_i(\boldsymbol{\xi})$ of gas molecules at the i-th rotational quantum level. It represents an extension of the Wang Chang Uhlenbecks equation to the case of degeneracy of energy levels which takes place for rotational degrees of freedom in the absence of magnetic field. Factor $w_{ij}^{kl} = q_k q_l / q_j q_i$ appears here due to the degeneracy of levels, with q_i being the degeneracy number.

When the equation is solved numerically, the rotational quantum levels are replaced with the nodes of discretization for angular velocity of rotation. In the general case numerical levels will not coincide with quantum levels. As demonstrated in [18], this approach introduces an error of the order $o(\Delta\omega^2)$, where $\Delta\omega$ is the step in the angular velocity grid. The angular velocity corresponding to the i-th level is equal to $\omega_i = \Delta\omega(i + 0.5)$, and the degeneracy of the numerical level equals $q_i = \omega_i$, $w_{ij}^{kl} = \omega_k\omega_l/\omega_j\omega_i$.

Similarly to the cases of a simple gas and a gas mixture, due to its symmetry properties, the collision integral of the generalized Boltzmann equation can be reduced to the following sum:

$$I_n(\boldsymbol{\xi}_\gamma) = \sum_{\nu=1}^{N_\nu} B_\nu \left(\delta_{n,k_\nu}\delta(\xi' - \xi_\gamma) + \delta_{n,l_\nu}\delta(\xi_1' - \xi_\gamma) + \delta_{n,i_\nu}\delta(\xi_{\alpha_\nu} - \xi_\gamma) - \delta_{n,j_\nu}\delta(\xi_{\beta_\nu} - \xi_\gamma)\right) \times$$
$$\times \Big(w_{ij}^{kl} f_l(\boldsymbol{\xi}_1') f_k(\boldsymbol{\xi}') - f_j(\boldsymbol{\xi}_{\beta_\nu}) f_i(\boldsymbol{\xi}_{\alpha_\nu})\Big) \left|\boldsymbol{\xi}_{\alpha_\nu} - \boldsymbol{\xi}_{\beta_\nu}\right|$$

Here, like in (3), $\delta(\xi_{\alpha_\nu} - \xi_\gamma) = \delta_{\alpha_\nu,\gamma}$, $\delta(\xi_{\beta_\nu} - \xi_\gamma) = \delta_{\beta_\nu,\gamma}$ are Kronecker symbols, and $f_i(\boldsymbol{\xi}_{\alpha_\nu}) = f_{i,\alpha_\nu}$, $f_j(\boldsymbol{\xi}_{\beta_\nu}) = f_{j,\beta_\nu}$ are the values of the grid distribution function. The Kronecker symbols δ_{n,i_ν}, δ_{n,j_ν}, δ_{n,k_ν}, δ_{n,l_ν} appear here because the sums form (11) are included in the cumulative sum over ν. In this case the integrating grid is 12-dimensional $\boldsymbol{\xi}_{\alpha_\nu}$, $\boldsymbol{\xi}_{\beta_\nu}$, $\boldsymbol{n}_\nu$, i_ν, j_ν, k_ν, l_ν.

The values of $\delta(\xi' - \xi_\gamma)$ and $\delta(\xi_1' - \xi_\gamma)$ are determined from formulae (1), and those of $f_l(\boldsymbol{\xi}_1')f_k(\boldsymbol{\xi}')$ from the formula (4). The conservation law similar to (2) for determination of the coefficient r_ν is in this case written as $\xi_{\alpha_\nu}^2 + I\omega_\alpha^2 + \xi_{\beta_\nu}^2 + I\omega_\beta^2 = (1 - r_\nu)(\xi_{\lambda_\nu}^2 + \xi_{\mu_\nu}^2) + r_\nu(\xi_{\lambda_\nu+s_\nu}^2 + \xi_{\mu_\nu-s_\nu}^2) + I\omega_\lambda^2 + I\omega_\mu^2$.

Calculation of differential collision cross-sections $\sigma_{ij}^{kl}(\theta, g)$ involves a considerable difficulty when a gas with rotational degrees of freedom is considered. As with the monatomic gas cross-sections, the cross-sections are pre-calculated, and the grid values are stored in data arrays. Here the scattering cross-sections are a function of 6 variables, and their storage calls for a large memory size (typically about 100MB). The cross-sections are calculated by the classical trajectory method. The simplest model of interaction between diatomic molecules is the rotator model. Each rotator possesses two centers of mass and two centers of forces which interaction is described by the Lennard-Jones potential: $U_{ij} = 4\varepsilon\left[(\sigma/r_{ij})^{12} - (\sigma/r_{ij})^{6}\right]$ where r_{ij} is the separation distance between the chosen pair of interacting centers. The rotators moment of inertia is $I = mr_a^2$, where r_a is the spacing between points where the mass resides.

Equations of interaction between two rotators have the form:

$$\frac{dX}{dt} = G(X), X = \begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \\ \boldsymbol{V} \\ \boldsymbol{\omega}_1 \\ \boldsymbol{\omega}_2 \end{pmatrix}, G = \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{\omega}_1 \times \boldsymbol{r}_1 \\ \boldsymbol{\omega}_2 \times \boldsymbol{r}_2 \\ \boldsymbol{F}/\mu \\ \boldsymbol{M}_1/I \\ \boldsymbol{M}_2/I \end{pmatrix}, \quad \begin{matrix} \boldsymbol{F} = \boldsymbol{F}_{11} + \boldsymbol{F}_{12} + \boldsymbol{F}_{21} + \boldsymbol{F}_{22} \\ \boldsymbol{M}_1 = \boldsymbol{r}_1 \times (-\boldsymbol{F}_{11} + \boldsymbol{F}_{12} - \boldsymbol{F}_{21} + \boldsymbol{F}_{22}) \\ \boldsymbol{M}_2 = \boldsymbol{r}_2 \times (\boldsymbol{F}_{11} + \boldsymbol{F}_{12} - \boldsymbol{F}_{21} - \boldsymbol{F}_{22}) \end{matrix} \tag{10}$$

Here $\boldsymbol{r}$ is the vector connecting the molecule centers, $\pm\boldsymbol{r}_{1,2}$ are relative mass position vectors, $\boldsymbol{\omega}_{1,2}$ are angular velocities of rotation of molecules, μ is the reduced molecule mass, $\boldsymbol{M}_{1,2}$ are force momenta, with values having subscripts 1 or 2 corresponding to the first and second molecule, respectively, and $\boldsymbol{F}_{1,2}$ is the force of interaction between the i-th center of the first molecule and the j-th center of the second molecule (i, j = 1,2).

Equations (10) are solved by Dormand-Prince method representing the 8th-order Runge-Kutta method. The calculations are done in dimensionless variables where σ, ε, $\mu/2$ are taken as the units of length, energy and mass. Fig. 2 below shows an example of the cross-section $\sigma_{ij}^{kl}(\theta, g)$ at $i = 10$, $j = 5$, $\theta = \pi/2$, $g = 5$, $\Delta\omega = 3$ for molecules of nitrogen (N_2) having $\sigma = 3.31A$, $\varepsilon = 37.3K$.

3 Examples of calculations

3.1 Heat transfer problem

The study of distribution of gas density, temperature and heat flux between two parallel plates having different temperatures is a fundamental problem of rarefied gas dynamics. Numerous works are devoted to this problem. In [4] the problem was solved for the hard-sphere potential by computing the collision integral with the use of a polynomial approximation of the distribution function.

Here the problem of heat transfer is studied by the projection method described in Section 2.1 above. The calculations are performed for a single-component monatomic gas. The cross-sections $\sigma_{LJ}(\theta, g)$ had been calculated in advance. There was used the Lennard-Jones potential which parameters for argon are $\varepsilon^{Ar} = 124K$, $\sigma^{Ar} = 3.418A$. The molecular mean free path was found using the omega integral value $\Omega^{(2,2)} = 1.093$ from [19].

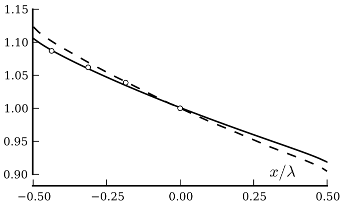


Figure 3: Density between the plates for Kn = 0.0658.

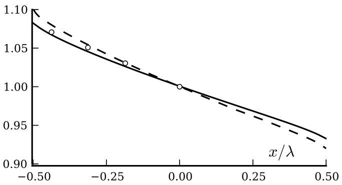


Figure 4: Density between the plates for Kn = 0.1395.

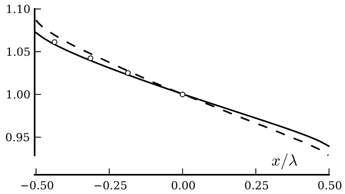


Figure 5: Density between the plates for Kn = 0.1942.

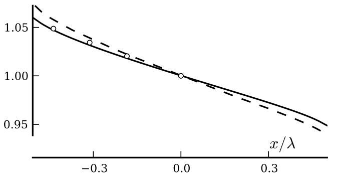


Figure 6: Density between the plates for Kn = 0.2992.

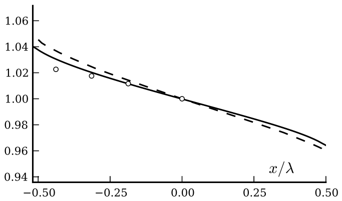


Figure 7: Density between the plates for Kn = 0.7582.

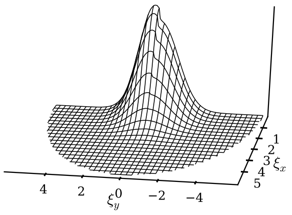


Figure 8: Distribution function at $x = 0$ for Kn = 0.7582.

Heat transfer between plates with temperatures $T_{\pm} = T(1 \pm \Delta T)$, $\Delta T = 0.14$ is considered. Diffuse-reflection boundary conditions with the accommodation coefficient $\alpha = 0.826$ are set on the plate walls. The problem statement models the experiments described in [20]. The heat transfer is considered for the following Knudsen numbers (Kn $= \lambda/d$, λ is the molecular mean free path, d is a distance between the plates): Kn $= 0.0658, 0.1395, 0.1942, 0.2992, 0.7582$. Results of the calculations are shown in Figs. 3, 4, 5, 6 and 7, respectively. The solid line corresponds to gas density distribution obtained with the projection method, the dotted line is plotted for results taken from [4] for the hard-sphere potential, and points show experimental data for argon from [20]. The comparison shows a good agreement with experimental data except for density in the vicinity of the cold plate for Kn $= 0.7582$ (see Fig. 7). It is interesting to note that in this case the velocity distribution function has a high gradient region on the plane $\xi_x = 0$ shown in Fig. 8. It is apparent that in this case the distribution is close to that composed by half-Maxwellians for temperatures $T_{\pm} = T(1 \pm \Delta T)$ that occurs in the free-molecular flow.

3.2 Shock wave structure

Determination of the shock wave structure is another classical problem of the kinetic theory. This problem was the subject of a number of experimental works and numerical simulations. Sufficiently reliable experimental data for monatomic gas Ar and diatomic gas N2 are

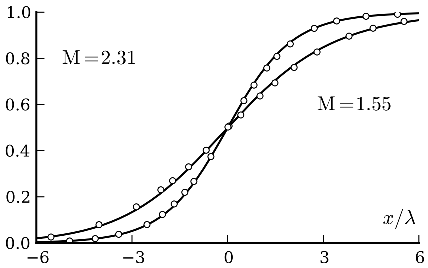


Figure 9: Density in the shock wave in the monatomic gas, ∘ – experimental data [21].

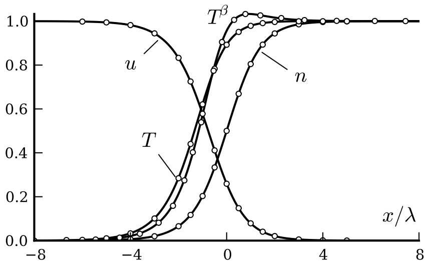


Figure 10: Structure of the shock wave in gas mixture for M = 3, $m^{\alpha}/m^{\beta} = 1/2$, $\chi^{\beta} = 0.1$ in comparison with data provided in [3].

published in [21]. There are a lot of publications on studies carried out using the statistical simulation method (DSMC) [1], [22].

Solution of the Boltzmann equation in case of the hard-sphere potential for a single-component gas and a gas mixture is presented in [2] and [3], respectively. There the collision integral is calculated using the same method of polynomial approximation of the distribution function as in [4]. By the projection method, the problem was studied, for example in [16].

In the present work the main part of the calculations is done for the realistic Lennard-Jones potential. Fig. 9 shows gas density profiles for shock waves with Mach numbers M = 1.55 and M = 2.31. Points correspond to the experimental data [21]. The solid line is plotted from calculations. Parameters of the Lennard-Jones potential for argon that were used are listed in the previous section.

Figs. 10 and 11 present results of calculations for the gas mixture. Fig. 10 compares our data obtained with the multipoint modification of the projection method (see Section 2.2

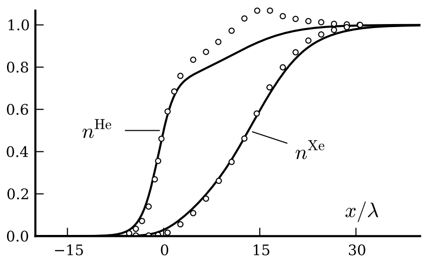


Figure 11: Density in the shock wave for M = 3.89 in the mixture of 97% helium and 3% xenon compared with experimental data [23].

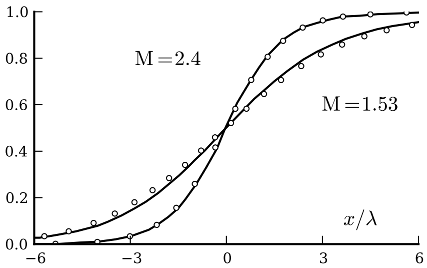


Figure 12: Density in the shock wave in the diatomic gas compared with experimental data [21].

above) with calculations presented in [3]. The hard-sphere potential was used. A comparison with experimental data [23] was also made (see Fig. 11). There was considered the case of the helium mixture with low molar fraction of xenon. The calculation was done for the following parameters of the Lennard-Jones potential: $\varepsilon^{He} = 10.2K$, $\varepsilon^{Xe} = 229K$, $\sigma^{He} = 2.576A$, $\sigma^{Xe} = 4.055A$ [19]. Parameters of interaction between different molecules were obtained on the basis of combination relations: $\varepsilon^{He,Xe} = \sqrt{\varepsilon^{He}\varepsilon^{Xe}}$, $\sigma^{He,Xe} = (\sigma^{He}\sigma^{Xe})/2$.

Fig. 12 shows results of the calculations for the diatomic gas performed through solution of the generalized Boltzmann equation by the method described in Section 2.3 above. The cross-sections $\sigma_{ij}^{kl}(\theta, g)$ were obtained with the classical trajectory method. A comparison with experimental data for argon taken from [21] is presented.

The comparisons show that for the monatomic gas the application of the Lennard-Jones potential gives results close to the experimental data. For the gas mixture, there is observed a good agreement in the density profile width (see Fig. 11). The peak in the helium profile may be associated with the experimental uncertainty. For the diatomic gas the agreement between the data is slightly worse. Probably, the model of interaction between diatomic molecules should be improved by introducing a dipole-dipole interaction and moving apart the repulsion and attraction centers [24].

Conclusion

The conservative method of calculating the collision integral for simple gases, gas mixtures and gases with rotational degrees of freedom of molecules is presented. It uses the common approach based on the projection method of summing up the contributions in the collision integral. For inert gases, the method allows calculating the integral for any molecular potentials. Gas mixtures can be effciently calculated for any molecular mass ratios and any concentrations of mixture components. For diatomic gases, there was developed the methodology of pre-calculating molecular collision cross-sections with translational-rotational energy transitions, which allows calculating the collision integral without the use of approximate models of inelastic collison. The algorithm of collision integral calculation was optimized, which provided for an essencial increase in the computation speed. Some sample of calculations presented in the paper confirm reliability and high accuracy of the obtained results, thus the efficiency of presented method for solving the Boltzmann equation.

References

[1] Bird, G. A. Oxford University Press, USA, 2nd edition, 1994.

[2] Ohwada, Taku. Physics of Fluids A: Fluid Dynamics 1993, 5(1):217–234.

[3] S. Kosuge, K. Aoki, S. Takata. European Journal of Mechanics, B/Fluids 2001, 20(1):87–126.

[4] Taku Ohwada. Phys. Fluids 1996; 8:2153.

[5] Palczewski, A., Schneider, J. and Bobylev, A. V. SIAM J. Numer. Anal. 1997; 34(5):1865–1883.

[6] Ibragimov, I. and Rjasanow, S. Computing 2002; 69(2):163–186.

[7] Mouhot, C. and Pareschi, L. Math. of Comp. 2006; 75:1833–1852.

[8] Gamba, I. M. and Tharkabhushanam, S. H. J. of Comput. Math. 2010.

[9] Cheremisin, F. G. Physics – Doklady 1997; 42:607–610.

[10] Tcheremissine, F. G. Comp. Math. and Math. Phys. 2006; 46(2):315–329.

[11] Cheremisin, F. G. Doklady Physics 2000; 45(8):401–404.

[12] F. G. Tcheremissine. Comp. Math. and Math. Phys. 2012; 52(2), 252–268.

[13] A.E. Beylich. Aachen: Technisch Hochcshule Report 2000.

[14] N. M. Korobov. Dokl. Akad. Nauk SSSR 1959; 124:1207–1210.

[15] Yu. A. Anikin, O. I. Dodulad, Yu. Yu. Kloss, D. V. Martynov, P. V. Shuvalov, F. G. Tcheremissine. Vacuum 2012; 86(11):1770–1777.

[16] Raines, A. A. AIP Conf. Proc. 2003; 663(1):67–76.

[17] O. I. Dodulad, F. G. Tcheremissine. In: 28th International Symposium on Rarefied Gas Dynamics 2012, AIP Conf. Proc. 1501, p. 302–309.

[18] Yu. A. Anikin, O. I. Dodulad. Comp. Math. and Math. Phys. 2013 (to be published)

[19] J. O. Hirschfelder, C. F. Curtiss, R. B. Bird. Wiley: University of Wisconsin, 1964.

[20] W. P. Teagan and G. S. Springer. Phys. Fluids 11, 497, 1968.

[21] H. Alsmeyer. J. Fluid. Mech 1976; 74:497–513.

[22] K. Koura. Phys. Fluids 9, 3543, 1997.

[23] A. S. Gmurczyk, M. Tarczynski, Z. A. Walenta. In: 11th International Symposium on Rarefied Gas Dynamics 1978; 1, p. 333–341.

[24] R. M. Berns and A. van der Avoird. J.Chem.Phys. 1980, 72:6107.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Finite-volume discretization of a 1d reaction-diffusion based multiphysics modelling of charge trapping in an insulator submitted to an electron beam irradiation

A. Aoufi[1] **and G. Damamme**[2]

[1] *Centre SMS, Laboratoire Georges Friedel, UMR CNRS 5307, Ecole des Mines de Saint-Etienne, 42023 Saint-Etienne Cedex.,*

[2] *CEA Gramat, BP 80200 - 46500 Gramat.,*

emails: `aoufi@emse.fr`, `gilles.damamme@cea.fr`

Abstract

This paper describes a new one-dimensional reaction-diffusion based modelling for charge trapping in an insulator submitted to an electron beam irradiation and its fully implicit finite-volume discretization.

Key words: reaction-diffusion equations, multiphysics coupling, charge trapping, implicit finite-volume scheme.

1 Introduction

Dielectric breakdown in insulating materials is of practical importance since it damages electronic devices [1]. Secondary electron emission yield (denoted by see) is one of the key parameters for dielectric materials and is the driving parameter of electric charging which can lead to electric breakdown. To study this phenomena, the behaviour of an insulator submitted to electron beam irradiation is considered. Several modelling based upon a two-fluxes method [2] were presented and discussed recently [3],[4],[5],[6],[7]. The purpose of this paper is to analyse a new mathematical modelling that is a reformulation in a reaction-diffusion framework of a modelling presented in [3], and is organized as follows. Section two presents the governing equations of the modelling which couples the trapped charge density $\rho(z,t)$, the electric field $E(z,t)$, the number of electron $n_e(z,t)$ or holes $n_h(z,t)$. Section three describes the fully implicit finite-volume numerical discretization scheme on a staggered refined mesh. A conclusion summarizes the paper.

2 Mathematical modelling

The purpose of the modelling is to analyze by numerical simulation the evolution of the global trapped charge per surface unit at time t, $Q_p(t)$ defined from the trapped charge $\rho(z,t)$

$$Q_p(t) = \int_\Omega \rho(z,t)\,dz \tag{1}$$

as a function of the true secondary electronic emission yield $see^*(t)$ defined by

$$see^*(t) = \kappa \frac{j_{e-}(0,t)}{j_0} \tag{2}$$

- The governing equation for the electric field $E(z,t)$ V/m is given by

$$\begin{aligned} \nabla.E(z,t) &= E_{ext} - \frac{\rho(z,t)}{\epsilon_0\epsilon_r} && (3)\\ E(0,t) &= \frac{Q_p(t)}{\epsilon_0\epsilon_r\,(1+\epsilon_r)} && (4) \end{aligned}$$

  where E_{ext} is the extracting field, ϵ_0 is the vacuum permittivity and ϵ_r is the material permittivity.

- The governing equation for the number of free charges $(n_c)_{c\in\{e,h\}}$ depends on the charge carrier flux $j_c(z,t)$ for $c \in \{e,h\}$ that is given by

$$j_c(z,t) = -D_c(E)\frac{\partial n_e}{\partial z} + \mu_c(E)v_c\,n_c(z,t) \tag{5}$$

  which is the sum of a diffusion term and a convection term. It is worth mentioning that the convection term can be positive or negative depending on the sign of the "mobility" coefficient μ_c. The conservation law fulfilled by $n_c(z,t)$ is therefore

$$\begin{aligned} &\frac{\partial n_c}{\partial t} + \frac{\partial}{\partial z}\left(-D_c(E)\frac{\partial n_e}{\partial z} + \mu_c(E)v_c\,n_c(z,t)\right) && (6)\\ =\;& 2S_c(z) - \frac{\sigma^+_{abs\cdot c} + \sigma^-_{abs\cdot c}}{2}\,n_c(z,t)\,v_c + j_c(z,t)\frac{\sigma^+_{abs\cdot c} - \sigma^-_{abs\cdot c}}{2} \end{aligned}$$

- The governing equation for the trapped charge density $\rho(z,t)$ is given by

$$\frac{\partial \rho(z,t)}{\partial t} + \nabla.j_T(z,t) = 0. \tag{7}$$

  The expression of $j_T(z,t)$ is given by

$$j_T(z,t) = j_h(z,t) - j_e(z,t) - j_0\,j_{prim}(z) \tag{8}$$

- The governing equation for the total trapped charge $Q_p(t)$ in C is related to the trapped charge density $\rho(z,t)$ thanks to the relation $Q_p(t) = \displaystyle\int_\Omega \rho(z,t)\,dz$. The integration of Eq.(7) over Ω leads to the relation

$$\frac{dQ_p(t)}{dt} + j_T(L,t) - j_T(0,t) = 0. \tag{9}$$

Taking into account the definition of the true secondary electron emission yield $see^*(t)$ given by

$$see^*(t) = v_e \frac{\kappa}{2-\kappa}\frac{n_e(0,t)}{j_0}. \tag{10}$$

leads to

$$\frac{dQ_p(t)}{dt} = j_0\,(1 - see^*(t)) + v_h\,n_h(L,t) - v_e\,n_e(L,t). \tag{11}$$

3 Numerical discretization scheme

The computational domain $\Omega = [0,L]$ is split into a set of I control volumes $\Omega_i = [z_i, z_{i+1}]$. The length of control volume Ω_i is m_i, its center is denoted by c_i. $d_{i,i+1}$ is the distance between the center of two adjacent control volumes Ω_i and Ω_{i+1}. The discrete unknowns located at the center of each control volume Ω_i are $n_e|_i^{k+1}$ and $n_h|_i^{k+1}$. The discrete unknowns located at the edge of each control volume Ω_i are ρ_i^{k+1}, E_i^{k+1} and $j_T|_i^{k+1}$. The time-discrete unknowns such as $Q_p|^{k+1}$ and $see^*|^{k+1}$. A backward Euler scheme is used to approximate the time-derivative and the integration of each governing equation over control volume Ω_i is given below.

- Discretization of the electric field equation leads to the discrete equation

$$E_{i+1}^{k+1} - E_i^{k+1} = \frac{m_i}{2}\frac{\rho_i^{k+1} + \rho_{i+1}^{k+1}}{\epsilon_0\epsilon_r} \tag{12}$$

where, a trapezoidal rule has been used to discretize the integral of the right hand side of Eq.(3). The discrete approximation of boundary condition Eq.(4) is written

$$E_1^{k+1} = E_{ext} - \frac{1}{\epsilon_0\epsilon_r\,(1+\epsilon_r)} Q_p^{k+1} \tag{13}$$

- Discretization of the charge density equation leads to the discrete equation

$$m_i.\frac{\rho_i^{k+1} - \rho_i^k}{\Delta t} + j_T|_{i+1}^{k+1} - j_T|_i^{k+1} = 0 \tag{14}$$

- Discretization of the number of free charge carrier's equation leads to the discrete equation

$$
\begin{aligned}
& m_i \frac{n_c|_i^{k+1} - n_c|_i^k}{\Delta t} - D_{i+\frac{1}{2}} \frac{n_c|_{i+1}^{k+1} - n_c|_i^{k+1}}{d_{i,i+1}} + D_{i-\frac{1}{2}} \frac{n_c|_i^{k+1} - n_c|_{i-1}^{k+1}}{d_{i-1,i}} \\
= \; & +2\, m_i\, S_c(c_i) - m_i \frac{\sigma_{abs{,}c}^{+,k+1} + \sigma_{abs{,}c}^{-,k+1}}{2}\, n_c|_i^{k+1}\, v_c \\
& + m_i\ \mu_c(E)|_i^{k+1}\, v_c\ n_c|_i^{k+1}\, \frac{\sigma_{abs{,}c}^{+,k+1} - \sigma_{abs{,}c}^{-,k+1}}{2}
\end{aligned}
\tag{15}
$$

Here $D_{i\pm\frac{1}{2}}$ is computed by the harmonic mean formula given by [8] in order to obtain the flux continuity across the interface between two adjacent control volumes. Adapted discrete equations are also written when the control volume is Ω_1 or Ω_I in order to take into account the Neuman type boundary conditions. A tridiagonal matrix has to be inverted by the forward/backward method, i.e. the (α, β)-TDMA algorithm given by [8] at each step of the nonlinear solver to compute approximation of $n_c|^{k+1,p+1}$ as a function of $n_c|^{k+1,p}$, where p is the nonlinear iteration index. This method is known to be stable if the matrix is diagonally dominant [8]. This is indeed the case and was effectively proved formally.

- Discretization of the true secondary electron emission yield $see^*(t)$ leads to the discrete equation

$$
see^*|^{k+1} = v_e \frac{\kappa}{2-\kappa} \frac{n_e|_1^{k+1}}{j_0}. \tag{16}
$$

- Discretization of the trapped charge density equation leads to the discrete equation

$$
\frac{Q_p|^{k+1} - Q_p|^k}{\Delta t} = j_0 \left(1 - see^*|^{k+1}\right) + v_h\, n_h|_I^{k+1} - v_e\, n_e|_I^{k+1}\,. \tag{17}
$$

4 Conclusion

In this paper we have presented a new 1D multiphysics modelling charge trapping in an insulator submitted to an electron beam irradiation. The validation of the fully implicit finite volume discretization presented in this paper is in progress.

References

[1] G. Damamme, C. Legressus, A.S. De Reggi, IEEE Trans.Dielec.Elect.Insul. 4 (5), 558-584, 1997

[2] S. Chandrasekhar (1960) Radiative transfer, éd New York: Dover, (ISBN 0486-6059-06).

[3] A.Aoufi, G. Damamme, "Analysis and Numerical Simulation of Secondary Electron Emission Yield of an Insulator submitted to An Electron Beam". Proceedings of 23rd International Symposium on Discharges and Electrical Insulation in Vacuum, Vols 1 and 2, pp.21-24, 2008.

[4] A.Aoufi, G.Damamme, "Numerical Computation of secondary electron emission yield by a two fluxes method", Numerical Analysis and Applied Mathematics, Vols 1-2, AIP Conference Proceedings, Vol 1168, pp 212-215, ISBN 978-0-7354-0709-1, International Conference on Numerical Analysis and Applied Mathematics, sep 18-22, Rethymno, Greece, 2009, Ed. Simos, TE; Psihoyios, G; Tsitouras, C.

[5] A.Aoufi, G. Damamme, "Numerical Simulation of Secondary Electron Emission Yield in the Case of Volume and Surface Trapping Evolution", Proceedings of 24th International Symposium on Discharges and Electrical Insulation in Vacuum, Vols 1 and 2, pp. 11-14, 2010.

[6] A.Aoufi, G.Damamme, "Two-Fluxes and Reaction-Diffusion Computation of Initial and Transient Secondary Electron Emission Yield by a Finite Volume Method", pp.89-108, in Numerical Simulations- Applications, Examples and Theory, Edited by Prof. Lutz Angerman, ISBN 978-953-307-440-5, INTECH open Editor, January 2011.

[7] A. Aoufi, G.Damamme, "1D numerical simulation of charge trapping in an insulator submitted to an electron beam irradiation. Part I: Computation of the initial secondary electron emission yield", Applied Mathematical Modelling, Vol. 35-3,pp 1175-1183.

[8] S.V. Patankar, Numerical Heat Transfer and Fluid Flow", McGRAW-HILL BOOK COMPANY, 1980.

*Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.*

On Soft Functions

Çiğdem Gündüz Aras[1], **Ayşe Sönmez**[2] **and Hüseyin Çakallı**[3]

[1] *Department of Mathematics, Kocaeli University*

[2] *Department of Mathematics, Gebze Institute of Technology*

[3] *Marmara Eğitim Köyü, Maltepe University*

emails: carasgunduz@gmail.com, asonmez@gyte.edu.tr, hcakalli@maltepe.edu.tr

Abstract

In this paper, we introduce soft continuous mappings which are defined over an initial universe set with a fixed set of parameters. Later we study soft open and soft closed mappings, soft homeomorphism and investigate some properties of these concepts.

Key words: soft sets; soft topology; soft continuity

1 Introduction

Most of the real life problems in social sciences, engineering, medical sciences, economics etc. the data involved are imprecise in nature. The solutions of such problems involve the use of mathematical principles based on uncertainty and imprecision. Thus classical set theory, which is based on the crisp and exact case may not be fully suitable for handling such problems of uncertainty. A number of theories have been proposed for dealing with uncertainties efficiently way. Some of these are theory of fuzzy sets [17], theory of intuitionistic fuzzy sets [3], theory of vague sets, theory of interval mathematics [4] ,[7] and theory of rough sets [14]. However, these theories have their own difficulties. Molodtsov [10] initiated a novel concept of soft sets theory as a new mathematical tool for dealing with uncertainties which is free from the above limitations. A soft set is a collection of approximate descriptions of an object. Soft systems provide a very general framework with the involvement of parameters. Since soft set theory has a rich potential, research on soft set theory and its applications in various fields are progressing rapidly. Maji et al. [11], [12] worked on soft set theory and presented an application of soft sets in decision making problems.

M.Shabir and M.Naz [16] introduced soft topological spaces. They also defined some concepts of soft sets on soft topological spaces. Later, researches about soft topological spaces were studied in [18]-[20]. In these studies, the concept of soft point is expressed by different approaches. In this paper we refer to the concept of soft point which was given in [22] .

The purpose of this paper is to investigate soft continuity on soft topological spaces.

2 Preliminaries

In this section, we give definitions and some results of soft sets.

Definition 1 *([10]) Let X be an initial universe set and E be a set of parameters. A pair (F, E) is called a soft set over X if only if F is a mapping from E into the set of all subsets of the set X, i.e., $F : E \to P(X)$, where $P(X)$ is the power set of X.*

Definition 2 *([11]) A soft set (F, A) over X is said to be a null soft set denoted by Φ if for all $a \in A$, $F(a) = \emptyset$ (null set).*

Definition 3 *([11]) A soft set (F, A) over X is said to be an absolute soft set denoted by $\tilde{A}$ if for all $a \in A$, $F(a) = X$.*

Definition 4 *[21] Let (F, A) and (G, B) be two soft sets over X . Then (F, A) is called a soft subset of (G, B), denoted by $(F, A) \subset (G, B)$, if*
(i) $A \subset B$,
(ii) $F(a) \subset G(a)$ for each $a \in A$.

Definition 5 *([11]) The intersection of two soft sets (F, A) and (G, B) over X is the soft set (H, C), where $C = A \cap B$ and $\forall c \in C$, $H(c) = F(c) \cap G(c)$. This is denoted by $(F, A)\tilde{\cap}(G, B) = (H, C)$.*

Definition 6 *([11]) The union of two soft sets (F, A) and (G, B) over X is the soft set, where $C = A \cup B$ and $\forall c \in C$,*

$$H(\varepsilon) = \begin{cases} F(\varepsilon), \text{ if } \varepsilon \in A - B \\ G(\varepsilon), \text{ if } \varepsilon \in B - A \\ F(\varepsilon) \cup G(\varepsilon), \text{ if } \varepsilon \in A \cap B \end{cases} .$$

This relationship is denoted by $(F, A)\tilde{\cup}(G, B) = (H, C)$.

Definition 7 *([16]) The complement of a soft set (F, E) is denoted by $(F, E)^c$ and is defined by $(F, E)^c = (F^c, E)$ where $F^c : E \to P(X)$ is a mapping given by $F^c(\alpha) = X - F(\alpha)$ for all $\alpha \in E$.*

Definition 8 *([16]) Let τ be a collection of soft sets over X. Then τ is said to be a soft topology on X if*
(1) Φ, $\tilde{X}$ belong to τ,
(2) the union of any number of soft sets in τ belongs to τ,
(3) the intersection of any two soft sets in τ belongs to τ.
The triplet (X, τ, E) is called a soft topological space over X.

Definition 9 *([16]) Let (X, τ, E) be a soft topological space over X. A soft set (F, E) over X is said to be a soft closed in X, if its relative complement $(F, E)^c$ belongs to τ.*

Proposition 1 *([16]) Let (X, τ, E) be a soft topological space over X. Then the collection $\tau_\alpha = \{F(\alpha) | (F, E) \in \tau\}$ defines a topology on X for each $\alpha \in E$.*

Definition 10 *([16]) Let (X, τ, E) be a soft topological space over X and (F, E) be a soft set over X. Then the soft closure of (F, E), denoted by $\overline{(F, E)}$ is the intersection of all soft closed super sets of (F, E).*

Definition 11 *([16]) Let (X, τ, E) be a soft topological space over X and (F, E) be a soft set over X. Then we associate with (F, E) a soft set over X, denoted by $(\overline{F}, E)$ and defined as $\overline{F}(\alpha) = \overline{F(\alpha)}$, where $\overline{F(\alpha)}$ is the closure of $F(\alpha)$ in τ_α for each $\alpha \in E$.*

Proposition 2 *([16]) Let (X, τ, E) be a soft topological space over X and (F, E) be a soft set over X. Then $(\overline{F}, E) \subset \overline{(F, E)}$.*

Corollary 1 *([16]) Let (X, τ, E) be a soft topological space over X and (F, E) be a soft set over X. Then $(\overline{F}, E) = \overline{(F, E)}$ if and only if $(\overline{F}, E)^c \in \tau$.*

Definition 12 *([23]) Let (X, τ, E) be a soft topological space over X, (G, E) be a soft set over X and $x \in X$. Then x is said to be a soft interior point of (G, E), if there exists a soft open set (F, E) such that $x \in (F, E) \subset (G, E)$.*

Definition 13 *([23]) Let (X, τ, E) be a soft topological space over X, (G, E) be a soft set over X and $x \in X$. Then (G, E) is said to be a soft neighbourhood of x, if there exists a soft open set (F, E) such that $x \in (F, E) \subset (G, E)$.*

Definition 14 *([23]) Let (X, τ, E) be a soft topological space over X then soft interior of a soft set (F, E) over X is denoted by $(F, E)^\circ$ and is defined as the union of all soft open sets contained in (F, E).*

Thus $(F, E)^\circ$ is the largest soft open set contained in (F, E).

3 Soft Continuity

Definition 15 *([22]) Let (F,E) be a soft set over X, and $x \in X$. The soft set (F,E) is called a soft point, denoted by (x_e,E), if for the element $e \in E$, $F(e) = \{x\}$ and $F(e^c) = \emptyset$ for all $e^c \in E - \{e\}$. In this case, we say that (x_e,E) is a point of a soft set (F,E).*

Definition 16 *Let (X,τ,E) and (Y,τ^c,E) be two soft topological spaces, $f : (X,\tau,E) \to (Y,\tau^c,E)$ be a mapping. For each soft neighbourhood (H,E) of $(f(x)_e,E)$, if there exists a soft neighbourhood (F,E) of (x_e,E) such that $f((F,E)) \subset (H,E)$, then f is said to be soft continuous mapping at (x_e,E).*
If f is soft continuous for all (x_e,E), then f is called soft continuous on X.

Theorem 1 *Let (X,τ,E) and (Y,τ^c,E) be two soft topological spaces, $f : (X,\tau,E) \to (Y,\tau^c,E)$ be a mapping. Then the following conditions are equivalent:*
(1) $f : (X,\tau,E) \to (Y,\tau^c,E)$ is a soft continuous mapping,
(2) For each soft open set (G,E) over Y, $f^{-1}((G,E))$ is a soft open set over X,
(3) For each soft closed set (H,E) over Y, $f^{-1}((H,E))$ is a soft closed set over X,
(4) For each soft set (F,E) over X, $f(\overline{(F,E)}) \subset \overline{(f(F,E))}$,
(5) For each soft set (G,E) over Y, $\overline{(f^{-1}(G,E))} \subset f^{-1}(\overline{(G,E)})$,
(6) For each soft set (G,E) over Y, $f^{-1}((G,E)^\circ) \subset (f^{-1}(G,E))^\circ$.

Proof: (1) ⇒ (2) Let (G,E) be a soft open set over Y and $(x_e,E) \in f^{-1}(G,E)$ be an arbitrary soft point. Then $f(x_e,E) = (f(x)_e,E) \in (G,E)$. Since f is soft continuous mapping, there exists $(x_e,E) \in (F,E) \in \tau$ such that $f(F,E) \subset (G,E)$. This implies that $(x_e,E) \in (F,E) \subset f^{-1}(G,E)$, $f^{-1}((G,E))$ is a soft open set over X.

(2) ⇒ (1) Let (x_e,E) be a soft point and $(f(x)_e,E) \in (G,E)$ be an arbitrary soft neighbourhood. Then $(x_e,E) \in f^{-1}(G,E)$ is a soft neighbourhood and $f(f^{-1}(G,E)) \subset (G,E)$.

(3) ⇒ (4) Let (F,E) be a soft set over X. Since $(F,E) \subset f^{-1}(f(F,E))$ and $f(F,E) \subset \overline{(f(F,E))}$, we have $(F,E) \subset f^{-1}(f(F,E)) \subset f^{-1}\overline{(f(F,E))}$. By part (3), since $f^{-1}\overline{(f(F,E))}$ is a soft closed set over X, $\overline{(F,E)} \subset f^{-1}\overline{(f(F,E))}$. Thus $f(\overline{(F,E)}) \subset f(f^{-1}\overline{(f(F,E))}) \subset \overline{f(F,E)}$ is obtained.

(4) ⇒ (5) Let (G,E) be a soft set over Y and $f^{-1}(G,E) = (F,E)$. By part (4), we have $f(\overline{(F,E)}) = f(\overline{f^{-1}(G,E)}) \subset \overline{f(f^{-1}(G,E))} \subset \overline{(G,E)}$. Then $\overline{f^{-1}(G,E)} = \overline{(F,E)} \subset f^{-1}(f\overline{(F,E)}) \subset f^{-1}(\overline{(G,E)})$.

(5) ⇒ (6) Let (G,E) be a soft set over Y. Substituting $(G,E)^c$ for condition in (5). Then $\overline{f^{-1}((G,E)^c)} \subset f^{-1}(\overline{(G,E)^c})$. Since $(G,E)^\circ = (\overline{(G,E)^c})^c$, then we have $f^{-1}((G,E)^\circ) =$

$$f^{-1}((\overline{(G,E)^c})^c) = (f^{-1}(\overline{(G,E)^c}))^c \subset (\overline{f^{-1}((G,E)^c)})^c = (\overline{(f^{-1}(G,E))^c})^c = (f^{-1}(G,E))^{\circ}.$$

(6) ⇒ (2) Let (G,E) be a soft open set over Y. Then since $(f^{-1}(G,E))^{\circ} \subset f^{-1}(G,E) = f^{-1}((G,E)^{\circ}) \subset (f^{-1}(G,E))^{\circ}$, $(f^{-1}(G,E))^{\circ} = f^{-1}(G,E)$ is obtained. This implies that $f^{-1}(G,E)$ is a soft open set over X. □

Example 1. Let $X = \{h_1, h_2, h_3\}$, $E = \{e_1, e_2\}$ and $\tau = \{\Phi, \tilde{X}, (F_1, E), (F_2, E)\}$, $\tau^c = \{\Phi, \tilde{X}, (G_1, E), (G_2, E)\}$ be two soft topologies defined on X, where (F_1, E), (F_2, E), (G_1, E) and (G_2, E) are soft sets over X, defined as follows:

$$F_1(e_1) = \{h_1, h_2\}, \quad F_1(e_2) = \{h_3\}, \quad F_2(e_1) = X, \quad F_2(e_2) = \{h_3\},$$

and

$$G_1(e_1) = \{h_1\}, \quad G_1(e_2) = \{h_3\}, \quad G_2(e_1) = \{h_1, h_3\}, \quad G_2(e_2) = \{h_2, h_3\},$$

If we get the mapping $f : X \to X$ defined as

$$f(h_1) = f(h_2) = h_1, \; f(h_3) = h_3$$

then since $f^{-1}(G_1, E) = (F_1, E)$ and $f^{-1}(G_2, E) = (F_2, E)$, f is a soft continuous mapping.

Example 2. Let $X = \{h_1, h_2, h_3\}$, $E = \{e_1, e_2\}$ and $\tau = \{\Phi, \tilde{X}, (F_1, E), (F_2, E), (F_3, E), (F_4, E)\}$, $\tau^c = \{\Phi, \tilde{X}, (G_1, E), (G_2, E), (G_3, E), (G_4, E)\}$ be two soft topologies defined on X where $(F_1, E), (F_2, E), (F_3, E), (F_4, E), (G_1, E), (G_2, E), (G_3, E)$ and (G_4, E) are soft sets over X, defined as follows:

$$F_1(e_1) = \{h_2\}, \quad F_1(e_2) = \{h_1\}, \quad F_2(e_1) = \{h_2, h_3\}, \quad F_2(e_2) = \{h_1, h_2\},$$

$$F_3(e_1) = \{h_3\}, \quad F_3(e_2) = \{h_1, h_2\}, \quad F_4(e_1) = \emptyset, \quad F_4(e_2) = \{h_1\},$$

$$F_5(e_1) = X, \quad F_2(e_2) = \{h_1, h_2\}$$

and

$$G_1(e_1) = \{h_2\}, \quad G_1(e_2) = \{h_1\}, \quad G_2(e_1) = \{h_2, h_3\}, \quad G_2(e_2) = \{h_1, h_2\},$$

$$G_3(e_1) = \{h_1, h_2\}, \quad G_3(e_2) = X, \quad G_4(e_1) = \{h_2\}, \quad G_4(e_2) = \{h_1, h_2\},$$

Then (X, τ, E), (X, τ^c, E) are two soft topological spaces and $f = 1_X : X \to X$ is not soft continuous mapping.

Theorem 2 *If $f : (X, \tau, E) \to (Y, \tau^c, E)$ is a soft continuous mapping, then for each $\alpha \in E$, $f_\alpha : (X, \tau_\alpha) \to (Y, \tau^c_\alpha)$ is a continuous mapping.*

Proof: Let $U \in \tau^c_\alpha$. Then there exists a soft open set (G, E) over Y such that $U = G(\alpha)$. Since $f : (X, \tau, E) \to (Y, \tau^c, E)$ is a soft continuous mapping, $f^{-1}(G, E)$ is a soft open set over X and $f^{-1}(G, E)(\alpha) = f^{-1}(G(\alpha)) = f^{-1}(U)$ is an open set. This implies that f_α is a continuous mapping. □

Now we give an example to show that the converse of above theorem does not hold.

Example 3. Let $X = \{x_1, x_2, x_3\}$, $Y = \{y_1, y_2, y_3\}$ and $E = \{e_1, e_2\}$. Then $\tau = \{\Phi, \tilde{X}, (F_1, E), (F_2, E), (F_3, E), (F_4, E), (F_5, E)\}$ is a soft topological space over X and $\tau^c = \{\Phi, \tilde{Y}, (G_1, E), (G_2, E), (G_3, E)\}$ is a soft topological space over Y. Here $(F_1, E), (F_2, E), (F_3, E), (F_4, E), (F_5,$ are soft sets over X and $(G_1, E), (G_2, E), (G_3, E)$ are soft sets over Y, defined as follows:

$$F_1(e_1) = \{x_1\}, \quad F_1(e_2) = \{x_1, x_3\}, \quad F_2(e_1) = \{x_2\}, \;\; F_2(e_2) = \{x_1\},$$

$$F_3(e_1) = \{x_1, x_2\}, \;\; F_3(e_2) = \{x_1, x_3\}, \quad F_4(e_1) = \emptyset, \quad F_4(e_2) = \{x_1\}, \quad F_5(e_1) = \{x_1, x_2\}, \quad F_5(e_2) = X.$$

and

$$G_1(e_1) = Y, \quad G_1(e_2) = \{y_2\}, \quad G_2(e_1) = \{y_1\}, \quad G_2(e_2) = \{y_2\}, \quad G_3(e_1) = \{y_1, y_2\}, \quad G_3(e_2) = \{y_2\}.$$

If we get the mapping $f : X \to Y$ defined as

$$f(x_1) = y_2, \; f(x_2) = y_1, \; f(x_3) = y_3$$

then f is not a soft continuous mapping, since $f^{-1}(G_1) \notin \tau$, where $f^{-1}(G_1)(e_1) = X$, $f^{-1}(G_1)(e_2) = \{x_1\}$. Also, $f_{e_1} : (X, \tau_{e_1}) \to (Y, \tau^c_{e_1})$ and $f_{e_2} : (X, \tau_{e_2}) \to (Y, \tau^c_{e_2})$ are continuous mappings. Here

$$\tau_{e_1} = \{\emptyset, X, \{x_1\}, \{x_2\}, \{x_1, x_2\}\}, \quad \tau_{e_2} = \{\emptyset, X, \{x_1\}, \{x_1, x_3\}\}$$

and

$$\tau^c_{e_1} = \{\emptyset, Y, \{y_1\}, \{y_1, y_2\}\}, \quad \tau^c_{e_2} = \{\emptyset, Y, \{y_2\}\}.$$

Now, let us show that when the above theorem is true.

Theorem 3 *If $(\overline{F}, E)^c$ is a soft open set over X, for each soft set (F, E), then $f : (X, \tau, E) \to (Y, \tau^c, E)$ is a soft continuous mapping if and only if $f_\alpha : (X, \tau_\alpha) \to (Y, \tau^c_\alpha)$ is continuous mapping, for each $\alpha \in E$.*

Proof: Let $f_\alpha : (X, \tau_\alpha) \to (Y, \tau^c_\alpha)$ be a continuous mapping, for each $\alpha \in E$, and let (F, E) be an arbitrary soft set over X. Then $f_\alpha(\overline{F}(\alpha)) \subset \overline{f(F)}(\alpha)$ is satisfied, for each $\alpha \in E$. Since $(\overline{F}, E)^c \in \tau$, $(\overline{F}, E) = \overline{(F, E)}$ from Corollary 1. Thus $f(\overline{(F, E)}) \subset \overline{f((F, E))}$ is obtained. This implies that $f : (X, \tau, E) \to (Y, \tau^c, E)$ is a soft continuous mapping. □

Definition 17 *Let (X, τ, E) and (Y, τ^c, E) be two soft topological spaces, $f : X \to Y$ be a mapping.*
a) If the image $f((F, E))$ of each soft open set (F, E) over X is a soft open set in Y, then f is said to be a soft open mapping.
b) If the image $f((H, E))$ of each soft closed set (H, E) over X is a soft closed set in Y, then f is said to be a soft closed mapping.

Proposition 3 *If $f : (X, \tau, E) \to (Y, \tau^c, E)$ is soft open (closed), then for each $\alpha \in E$, $f_\alpha : (X, \tau_\alpha) \to (Y, \tau^c_\alpha)$ is an open (closed) mapping.*

Proof: The proof of the proposition is straightforward and it is left to the reader. □

Note that the concepts of soft continuous, soft open and soft closed mappings are all independent of each other.

Example 4. Let (X, τ, E) be soft discrete topological space and (X, τ^c, E) be soft indiscrete topological space. Then $1_X : (X, \tau, E) \to (X, \tau^c, E)$ is a soft open and soft closed mapping. But it is not soft continuous mapping.

Example 5. Let $X = \{h_1, h_2, h_3\}$, $E = \{e_1, e_2\}$ and $\tau = \{\Phi, \tilde{X}, (F_1, E), (F_2, E), ..., (F_7, E)\}$, $\tau^c = \{\Phi, \tilde{X}, (G_1, E), (G_2, E), (G_3, E), (G_4, E)\}$ be two soft topologies defined on X where $(F_1, E), (F_2, E), (F_3, E), ..., (F_7, E), (G_1, E), (G_2, E), (G_3, E)$ and (G_4, E) are soft sets over X, defined as follows:

$$F_1(e_1) = \{h_2\}, \quad F_1(e_2) = \{h_1\}, \quad F_2(e_1) = \{h_1, h_3\}, \quad F_2(e_2) = \{h_2, h_3\},$$

$$F_3(e_1) = \{h_2\}, \quad F_3(e_2) = X, \quad F_4(e_1) = \emptyset, \quad F_4(e_2) = \{h_1\},$$

$$F_5(e_1) = \{h_1, h_3\}, \quad F_5(e_2) = X, \quad F_6(e_1) = \emptyset, \quad F_6(e_2) = \{h_2, h_3\},$$

$$F_7(e_1) = \emptyset, \quad F_7(e_2) = X$$

and

$$G_1(e_1) = \{h_2\}, \quad G_1(e_2) = \{h_1\}, \quad G_2(e_1) = \{h_2, h_3\}, \quad G_2(e_2) = \{h_1, h_2\},$$

$$G_3(e_1) = \{h_1, h_2\}, \quad G_3(e_2) = X, \quad G_4(e_1) = \{h_2\}, \quad G_4(e_2) = \{h_1, h_2\}.$$

If we get the mapping $f : X \to X$ defined as $f(h_i) = h_1$, for $1 \le i \le 3$. It is clear that

$$f^{-1}(G_1)(e_1) = f^{-1}(G_4)(e_1) = \emptyset, \;\; f^{-1}(G_1)(e_2) = f^{-1}(G_4)(e_2) = X, \;\; f^{-1}(G_3)(e_1) = X, \;\; f^{-1}(G_3)(e_2) = X.$$

Then f is a soft continuous mapping, but

$$f(F_1)(e_1) = \{h_1\}, \;\; f(F_1)(e_2) = \{h_1\}, \;\; f(F_1^c)(e_1) = \{h_1\}, \;\; f(F_1^c)(e_2) = \{h_1\}.$$

Hence it is not both soft open and soft closed mapping.

Example 6. Let $X = \{h_1, h_2, h_3\}, Y = \{a, b\}$ and $E = \{e_1, e_2\}$ and $\tau = \{\Phi, \tilde{X}, (F_1, E), (F_2, E)\}$, $\tau^c = \{\Phi, \tilde{Y}, (G_1, E), (G_2, E)\}$ be two soft topologies defined on X and Y, respectively. Here $(F_1, E), (F_2, E), (G_1, E), (G_2, E)$ are soft sets over X and Y, respectively. The soft sets are defined as follows:

$$F_1(e_1) = \{h_1, h_2\}, \quad F_1(e_2) = \{h_3\}, \quad F_2(e_1) = X, \quad F_2(e_2) = \{h_3\},$$

and

$$G_1(e_1) = Y, \quad G_1(e_2) = \{b\}, \quad G_2(e_1) = \{a\}, \quad G_2(e_2) = \{b\},$$

If we get the mapping $f : X \to Y$ defined as

$$f(h_1) = \{a\}, \quad f(h_2) = f(h_3) = \{b\}.$$

It is clear that

$$f(F_1)(e_1) = Y, \;\; f(F_1)(e_2) = \{b\}, \;\; f(F_2)(e_1) = Y, \;\; f(F_2)(e_2) = \{b\}.$$

Then the mapping $f : X \to Y$ is a soft open mapping. Also since $f(F_1^c)(e_1) = \{b\}, \;\; f(F_1^c)(e_2) = Y$, it is not soft closed mapping and $f^{-1}(G_1)(e_1) = X$, $f^{-1}(G_1)(e_2) = \{h_2, h_3\}$. Hence it is not soft continuous mapping.

Example 7. Let $X = \{h_1, h_2, h_3\}, Y = \{a, b\}, E = \{e_1, e_2\}$ and $\tau = \{\Phi, \tilde{X}, (F_1, E), (F_2, E), (F_3, E)\}$, $\tau^c = \{\Phi, \tilde{Y}, (G_1, E), (G_2, E)\}$ be two soft topologies defined on X and Y, respectively. Here $(F_1, E), (F_2, E), (F_3, E), (G_1, E), (G_2, E)$ are soft sets over X and Y, respectively. The soft sets are defined as follows:

$$F_1(e_1) = \{h_1, h_3\}, \; F_1(e_2) = \{h_2\}, \; F_2(e_1) = X, \; F_2(e_2) = \{h_2, h_3\}, \; F_3(e_1) = \{h_3\}, \; F_3(e_2) = \{h_2\}$$

and

$$G_1(e_1) = \emptyset, \quad G_1(e_2) = \{a\}, \quad G_2(e_1) = \{a\}, \quad G_2(e_2) = Y.$$

Now we define the mapping $f : X \to Y$ as $f(h_1) = f(h_2) = \{a\}$, $f(h_3) = \{b\}$. It is clear that $f(F_1^c(e_1)) = f(h_2) = \{a\}, f(F_1^c(e_2)) = f(\{h_1, h_3\}) = Y$, $f(F_2^c(e_1)) = \emptyset, f(F_2^c(e_2)) = f(\{h_1\}) = \{a\}$, $f(F_3^c(e_1)) = \{a\}, f(F_3^c(e_2)) = Y$. This implies that f is a soft closed mapping. Also $f(F_1(e_1)) = Y, f(F_1(e_2)) = \{a\}$, $f^{-1}(G_1(e_1)) = \emptyset, f^{-1}(G_1(e_2)) = \{h_1, h_2\}$. Then it is not soft open and soft continuous mapping, respectively.

Theorem 4 *Let (X, τ, E) and (Y, τ^c, E) be two soft topological spaces, $f : X \to Y$ be a mapping.*
a) f is a soft open mapping if and only if for each soft set (F, E) over X, $f((F, E)^\circ) \subset (f(F, E))^\circ$ is satisfied.
b) f is a soft closed mapping if and only if for each soft set (F, E) over X, $(\overline{f(F, E)}) \subset f(\overline{(F, E)})$ is satisfied.

ISBN: 978-84-616-2723-3

Proof: a) Let f be a soft open mapping and (F,E) be a soft set over X. $(F,E)^\circ$ is a soft open set and $(F,E)^\circ \subset (F,E)$. Since f is a soft open mapping, $f((F,E)^\circ)$ is a soft open set in Y and $f((F,E)^\circ) \subset f((F,E))$. Thus $f((F,E)^\circ) \subset f((F,E))^\circ$ is obtained.
Conversely, let (F,E) be any soft open set over X. Then $(F,E) = (F,E)^\circ$. From the condition of theorem, we have $f((F,E)^\circ) \subset (f(F,E))^\circ$. Then $f((F,E)) = f((F,E)^\circ) \subset (f(F,E))^\circ \subset f((F,E))$. This implies that $f((F,E)) = (f(F,E))^\circ$. This completes the proof.
b) Let f be a soft closed mapping and (F,E) be any soft set over X. Since f is a soft closed mapping, $f(\overline{(F,E)})$ is a soft closed set over Y and $f((F,E)) \subset f(\overline{(F,E)})$. Thus $\overline{f(F,E)} \subset f(\overline{(F,E)})$ is obtained.
Conversely, let (F,E) be any soft closed set over X. From the condition of theorem, $\overline{(f(F,E))} \subset f(\overline{(F,E)}) = f((F,E)) \subset \overline{(f(F,E))}$. This means that $\overline{(f(F,E))} = f((F,E))$. This completes the proof. □

Definition 18 *Let (X,τ,E) and (Y,τ^c,E) be two soft topological spaces, $f : X \to Y$ be a mapping. If f is a bijection, soft continuous and f^{-1} is a soft continuous mapping, then f is said to be soft homeomorphism from X to Y. When a homeomorphism f exists between X and Y, we say that X is soft homeomorphic to Y.*

Theorem 5 *Let (X,τ,E) and (Y,τ^c,E) be two soft topological spaces, $f : X \to Y$ be a bijective mapping. Then the following conditions are equivalent:*
(1) f is a soft homeomorphism,
(2) f is a soft continuous and soft closed mapping,
(3) f is a soft continuous and soft open mapping.

Proof: It is easily obtained. □

4 Conclusion

We have introduced soft continuous mappings which are defined over an initial universe with a fixed set of parameters. Later we study soft open and soft closed mappings, soft homeomorphism and investigate some properties of these concepts. In the end, we must say that, soft topological spaces are defined over different set of parameters and it does not make any sense for the results of this paper.

References

[1] M.I.Ali, F.Feng, X.Y.Liu, W.K.Min, M.Shabir, On some new operations in soft set theory, Comput. Math. Appl. 57 (2009) 1547-1553

[2] H.Aktas, N. agman, Soft sets and soft group, Information Science 177 (2007) 2726-2735.

[3] K.Atanassov, Intuitionistic fuzzy sets, Fuzzy Sets and Systems 20 (1986) 87-96.

[4] K.Atanassov, Operators over interval valued intuitionistic fuzzy sets, Fuzzy Sets and Systems 64 (1994) 159-174.

[5] D.Chen, The parametrization reduction of soft sets and its applications, Comput. Math. Appl. 49 (2005) 757-763

[6] F. Feng, Y.B. Jun, X. Zhao, Soft semirings, Comput. Math. Appl.56 (2008) 2621-2628.

[7] M.B.Gorzalzany, A method of inference in approximate reasoning based on interval-valued fuzzy sets, Fuzzy Sets and Systems 21 (1987) 1-17.

[8] Y.B.Jun, C.H.Park, Applications of soft sets in ideal theory of BCK/BCI-Algebras, Inform.Sci. 178 (2008) 2466-2475.

[9] O.Kazanc, .Ylmaz and S.Yamak, Soft sets and soft BCH-Algebras, Hacettepe Journal of Mathematics and Statistics 39(2) (2010) 205-217.

[10] D. Molodtsov, Soft set theory- first results, Comput. Math. Appl.37 (1999) 19-31.

[11] P.K.Maji, R.Bismas, A.R.Roy, Soft set theory, Comput. Math. Appl.45 (2003) 555-562.

[12] P.K.Maji, A.R.Roy, R.Bismas, An Application of soft sets in a decision making problem, Comput. Math. Appl.44 (2002) 1077-1083

[13] Qiu-Mei Sun, Zi-Liong Zhang, Jing Liu, Soft sets and soft modules, Lecture Notes in Comput. Sci. 5009 (2008) 403-409.

[14] Z.Pawlak, Rough sets, Int.J.Comput.Sci. 11 (1982) 341-356.

[15] M.Shabir, M.Irfan Ali, Soft ideals and generalized fuzzy ideals in semigroups, New Math. Nat. Comput. 5 (2009) 599-615.

[16] M.Shabir, M. Naz, On soft topological spaces, Comput. Math. Appl. 61 (2011) 1786-1799.

[17] L.A.Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338-353.

[18] I. Zorlutuna, M. Akdag, W.K. Min, S. Atmaca, Remarks on soft topological spaces, Annals of Fuzzy Mathematics and Informatics, 2012.

[19] N. Cagman, S. Karatas, S. Enginoglu, Soft topology, Comput. Math. Appl. 62 (2011) 351-358.

[20] W.K. Min, A note on soft topological spaces, Comput. Math. Appl. **62** (2011) 3524-3528.

[21] P. Majumdar, S.K. Samanta, On soft mappings, Comput. Math. Appl., **60** (2010) 2666-2672

[22] S. Bayramov, C. Gunduz(Aras), Soft locally compact and soft paracompact spaces, Journal of Mathematics and System Science

[23] S.Hussain, B.Ahmad, Some properties of soft topological spaces, Comput. Math. Appl.62 (2011) 4058-4067

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Numerical solution of time-dependent Maxwell's equations for modeling scattered electromagnetic wave's propagation

Adérito Araújo[1], Sílvia Barbeiro[1], Luís Pinto[1], Francisco Caramelo[2], António L. Correia[2], Miguel Morgado[2,5], Pedro Serranho[2,4], Ana Sílvia F. C. Silva[5] and Rui Bernardes[2,3]

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

[2] *IBILI, Faculty of Medicine, University of Coimbra, Portugal*

[3] *ABILI, Coimbra, Portugal*

[4] *Mathematics Section, Department of Science and Technology, Open University, Portugal*

[5] *Department of Physics, Faculty of Science and Technology, University of Coimbra, Portugal*

emails: alma@mat.uc.pt, silvia@mat.uc.pt, luisp@mat.uc.pt, fcaramelo@fmed.uc.pt, acorreia@aibili.pt, mmorgado@ibili.uc.pt, Pedro.Serranho@uab.pt, anascsilva@gmail.com, rmbernardes@fmed.uc.pt

Abstract

We present the discontinuous Galerkin method combined with a low-storage Runge-Kutta method as an accurate and efficient way to numerically solve the time-dependent Maxwell's equations. We investigate the numerical scheme in the context of modeling scattered electromagnetic wave's propagation through human eye's structures.

Key words: Maxwell's equations, discontinuous Galerkin method, low-storage Runge-Kutta method, anisotropic permittivity tensor

1 Introduction

The human retina is a complex structure in the eye that is responsible for the sense of vision. It is part of the central nervous system, constituted by layers of neurons interconnected through synapses [7]. There are ten layers in total; one of them is constituted by photosensitive neurons. There are a number of eye-related pathologies that can be identified by analysis of these retinal layers in detail [7]. All these pathologies can be diagnosed more conclusively with the help of the increasingly popular optical imaging technique – optical coherence tomography (OCT) [3], [15]. In fact, previous studies have established a link between changes in the blood-retina barrier and in optical properties of the retina [1], [2] which can be identified by this exam.

Physically, OCT it is based in low coherence interferometry. This technique uses an electromagnetic wave with a low coherence length. In order to better understand the information carried in an optical coherence tomography, it is crucial to study in detail the behaviour of the electromagnetic wave as it travels through the sample. Several different models have been developed to describe the interactions of the electromagnetic field with biological structures. The first models were based on single-scattering theory [14], which is restricted to superficial layers of highly scattering tissue in which only single scattering occurs. Simulating the full complexity of the retina, in particular the variation of the size and shape of each structure, distance between them and the respective refractive indexes, requires a more rigorous approach that can be achieved by solving Maxwell's equations.

In this work we discuss the numerical discretization of the time-dependent Maxwell's equations. We use the discontinuous Galerkin (DG) method for the integration in space and a low-storage Runge-Kutta method for the integration in time. In the model we consider anisotropic permittivity tensors which arise naturally in our application of interest. We illustrate the performance of the method with some numerical experiments.

2 Maxwell's Equations

We shall consider the time domain Maxwell's equations in the two-dimensional transverse electric (TE) mode. For this case, and assuming no conductivity effects, the equations in the non-dimensional form are

$$\epsilon \frac{\partial E^x}{\partial t} = \frac{\partial H^z}{\partial y} \tag{1}$$

$$\epsilon \frac{\partial E^y}{\partial t} = -\frac{\partial H^z}{\partial x} \tag{2}$$

$$\mu \frac{\partial H^z}{\partial t} = -\frac{\partial E^y}{\partial x} + \frac{\partial E^x}{\partial y}, \qquad \text{in } \Omega \times (0,T], \tag{3}$$

where $E = (E^x, E^y)$ and H^z represent the electric and magnetic fields, respectively, ϵ represents the relative permittivity of the medium and μ is the permeability of the medium and Ω is a two-dimensional domain. The nondimensionalized variables in (1)-(3) are related to the physical variables in the following way

$$\frac{\tilde{x}}{L} = x, \quad \frac{\tilde{y}}{L} = y, \quad \frac{c\tilde{t}}{L} = t, \quad E = Z_0^{-1} E, \quad H^z = H^z,$$

where L is a reference length, c is the speed of light in free space, and $Z_0 = \sqrt{\mu_0/\epsilon_0}$ is the free-space impedance. The set of equations (1)-(3) must be complemented by proper boundary conditions, for instance, the perfect electric boundary condition (PEC) [5]

$$\eta \times E = 0, \qquad \text{on } \partial\Omega, \tag{4}$$

or the Silver-Müller absorbing boundary condition [9]

$$\eta \times E = \sqrt{\frac{\mu}{\epsilon}}\eta \times (H^z \times \eta), \qquad \text{on } \partial\Omega. \tag{5}$$

In both cases η denotes the unit outward normal. To complete the model, initial conditions

$$E_0 = E(0) \quad \text{and} \quad H_0 = H(0), \qquad \text{in } \Omega,$$

must also be considered.

3 The scattered-field formulation

For the investigation of scattering problems in linear materials, we exploit the linearity of the Maxwell's equations (1)-(3) and separate the fields (E, H) into incident (E^i, H^i) and scattered components (E^s, H^s), i.e.,

$$E = E^s + E^i \qquad \text{and} \qquad H = H^s + H^i. \tag{6}$$

Assuming that the incident field is also a solution of the Maxwell's equations we obtain, after some manipulation the scattered field formulation as in [16],

$$\epsilon\frac{\partial E^{x,s}}{\partial t} = \frac{\partial H^{z,s}}{\partial y} + P \tag{7}$$

$$\epsilon\frac{\partial E^{y,s}}{\partial t} = -\frac{\partial H^{z,s}}{\partial x} + Q \tag{8}$$

$$\mu\frac{\partial H^{z,s}}{\partial t} = -\frac{\partial E^{y,s}}{\partial x} + \frac{\partial E^{x,s}}{\partial y} + R, \qquad \text{in } \Omega \times (0,T] \tag{9}$$

with the source terms

$$P(x,y,t) = (\epsilon^i - \epsilon)\frac{\partial E^{x,i}}{\partial t},$$

$$Q(x,y,t) = (\epsilon^i - \epsilon)\frac{\partial E^{y,i}}{\partial t},$$

$$R(x,y,t) = (\mu^i - \mu)\frac{\partial H^{z,i}}{\partial t},$$

where ϵ^i and μ^i represent the relative permittivity and permeability of the medium in which the incident field propagates. In our research, for simplicity, we adopt the pure scattered field formalism in opposition to the total field/scattered field formalism [16].

4 Numerical method

The DG method was first introduced in [11]. In particular, the nodal formulation described in [6] has gained notorious popularity in recent years and it has been extensively used in electromagnetic problems since the first application of the method to Maxwell's equations in 2002 (see [5]). This is the method that we have chosen to use. In opposition to the traditional finite difference (FD) time domain methods [16] based on the Yee's scheme [17] the DG method is a high-order accurate method that can easily handle complex geometries. Moreover, local refinement strategies can easily be incorporate due to the ability of the method to deal with irregular meshes with hanging nodes and local spaces of different orders. When compared to finite element methods [12], the DG method presents the advantages of avoiding the solution of linear systems when explicit time integrators are employed and the suitability for parallel implementation on modern multi-graphics processing units (GPUs) [6].

As starting point of our work we use the MatLab codes for the DG method [6]. The software includes many attributes that we intent to exploit in future work such as three-dimensional routines and nonconforming triangulations. For now, we are concerned with two-dimensional problems and we have restricted our attention to conform triangular elements. Our main focus has been on the implementation of additional features needed for the kind of problems which are in our objectives. For instance, the retinal nerve fiber layer of the retina is a birefringent medium, meaning that tensorial permittivity ϵ needs to be taken into account. We addressed this issue using the numerical scheme presented in [8]. The procedure, based on the so called upwind flux, is fairly simple and can be easily incorporated in the algorithm. Nevertheless, it seems to be efficient and robust for a wide range of problems, since the two by two matrix that represents the tensor ϵ only needs to be invertible. In [8] the authors claim that the method, which is an extension of the basic two-dimensional formulation for isotropic materials to allow anisotropic permittivity tensors, retains the convergence characteristics of the original method. The validation tests that we present in the next section corroborate those findings.

Another aspect of our study concerns the integration in time. The time integration scheme used in the original algorithm [6] is a fourth order, five stage low-storage Runge-Kutta (LSRK) method. Such type of schemes are very popular in this context as they retain the qualities of the original Runge-Kutta schemes while decreasing the memory consumption significatively. However, we recall that explicit methods like the LSRK scheme are subject to the CFL stability condition. In practice, this means that the time step size is proportional to the smallest elements of the spatial mesh [6]. For some problems this condition can have a strong impact on the efficiency of the algorithm. In the future we intent to deal with this issue using local time-stepping strategies [6] or locally implicit time-schemes [4]. For now, we mitigate this restriction implementing the improved fourth order, 14-stage LSRK presented in [10]. With this scheme the authors report a speed improvement of about $40\% - 50\%$ in relation to the previous fourth order, five stage LSRK method. This gain is consequence of an improved stability region possible by a suitable choice of the coefficients of the scheme. This is a very attractive approach, with very few changes in the code and no additional cost in memory or accuracy. The validation tests that we present in the next section indicate that these conclusions are also valid for the tensorial case.

An important aspect in computational electromagnetic problems is the implementation of absorbing boundary conditions. In our work we have implemented the Silver-Müller boundary condition (5) and the well established and more effective uniaxial perfectly matched layer (UPML) [13]. The UPML can be incorporated in the DG method without any major modification and so far has shown to be an efficient approach.

5 Numerical results

We consider initial conditions, boundary conditions and functions $P(t)$, $Q(t)$ and $R(t)$, such that the system of equations (7)-(9) has the solution

$$E^x = t^2(-\cos(\pi x) - 1)(\frac{1}{3}y^3 - y) \tag{10}$$

$$E^y = 0 \tag{11}$$

$$H^z = \frac{1}{3}t^3(-\cos(\pi x) - 1)(y^2 - 1) \tag{12}$$

with $\epsilon = 4x_c^2 + y_c^2 + 1$, for the results in Tables 1 and 2, and

$$\epsilon = \left[\begin{array}{cc} 4x_c^2 + y_c^2 + 1 & |x_c + y_c| \\ |x_c + y_c| & x_c^2 + 2 \end{array} \right]. \tag{13}$$

for the results in Tables 3-6. In both cases we set $\mu = 1$, $\Omega = [-1, 1]^2$ and $T = 5$. By x_c and y_c we represent the centroid of the triangles. Note that in the implementation ϵ must be constant in each triangle.

h_{max}	$\|E^x - E_h^x\|_{L^2}$	Rate	$\|E^y - E_h^y\|_{L^2}$	Rate	$\|H^z - H_h^z\|_{L^2}$	Rate
1.4142e-01	4.1644e-01	2.0580	4.1702e-01	2.0068	3.5908e-01	2.0297
7.0711e-02	2.4017e-02	2.0265	2.5820e-02	2.0002	2.1537e-02	2.0101
3.5355e-02	1.4469e-03	-	1.6133e-03	-	1.3273e-03	-

Table 1: Convergence rate in the L^2 norm for polynomial approximation of order 1.

From the analysis of Tables 1-4 we observe that the tensorial scheme preserves the optimal rate of convergence of the upwind flux in the scalar case $O(h^{N+1})$, where N denotes the order of the polynomial involved in the approximation. Note also that the order of the

h_{max}	$\|E^x - E^x_h\|_{L^2}$	Rate	$\|E^y - E^y_h\|_{L^2}$	Rate	$\|H^z - H^z_h\|_{L^2}$	Rate
1.4142e-01	8.6363e-04	3.0268	1.0196e-03	3.0124	1.7076e-03	2.9941
7.0711e-02	1.3002e-05	3.0120	1.5660e-05	3.0070	2.6901e-05	2.9978
3.5355e-02	1.9981e-07	-	2.4231e-07	-	4.2159e-07	-

Table 2: Convergence rate in the L^2 norm for polynomial approximation of order 2.

h_{max}	$\|E^x - E^x_h\|_{L^2}$	Rate	$\|E^y - E^y_h\|_{L^2}$	Rate	$\|H^z - H^z_h\|_{L^2}$	Rate
1.4142e-01	5.4515e-01	2.0461	6.2046e-01	1.9949	3.9240e-01	2.0202
7.0711e-02	3.1963e-02	2.0197	3.9055e-02	1.9964	2.3847e-02	2.0067
3.5355e-02	1.9438e-03	-	2.4532e-03	-	1.4767e-03	-

Table 3: Convergence rate in the L^2 norm for polynomial approximation of order 1.

h_{max}	$\|E^x - E^x_h\|_{L^2}$	Rate	$\|E^y - E^y_h\|_{L^2}$	Rate	$\|H^z - H^z_h\|_{L^2}$	Rate
1.4142e-01	8.1456e-04	2.9993	9.4071e-04	3.0184	1.7020e-03	2.9940
7.0711e-02	1.2739e-05	2.9996	1.4328e-05	3.0091	2.6817e-05	2.9979
3.5355e-02	1.9916e-07	-	2.2106e-07	-	4.2022e-07	-

Table 4: Convergence rate in the L^2 norm for polynomial approximation of order 2.

	h_{max}	$\|E_x - E^h_x\|_{L^2}$	Time
original fourth order, 5-stage LSRK	7.0711e-02	3.1963e-02	69.25s
improved fourth order, 14-stage LSRK	7.0711e-02	3.1963e-02	38.35s (44%)
original fourth order, 5-stage LSRK	3.5355e-02	1.9438e-03	514.26s
improved fourth order, 14-stage LSRK	3.5355e-02	1.9438e-03	295.99s (42%)

Table 5: Comparison of the two LSTR methods for polynomial approximation of order 1.

	h_{max}	$\|E_x - E^h_x\|_{L^2}$	Time
original fourth order, 5-stage LSRK	7.0711e-02	1.2739e-05	132.43s
improved fourth order, 14-stage LSRK	7.0711e-02	1.2739e-05	79.71s (39%)
original fourth order, 5-stage LSRK	3.5355e-02	1.9916e-07	1259.94s
improved fourth order, 14-stage LSRK	3.5355e-02	1.9916e-07	745.40s (40%)

Table 6: Comparison of the two LSTR methods for polynomial approximation of order 2.

errors is about the same when we compare the problem with tensorial permittivity and the problem with scalar permittivity.

In Tables 5 and 6 we compare the two different implementations for the time integration. The error is exactly the same for both methods and the running time decreases about 40%.

Even though we only show the results for Ex, in our experiments we found that the error of the other field components, E^y and H^z, present the same type of behaviour.

To illustrate the advantages of the DG method over the FD method we examine the scattering by a dielectric cylinder with a relative permittivity of $\epsilon = 5$. This scatterer with radius $r = 0.5$ and centered at the origin is excited by a Gaussian plane wave propagating along the y direction. We measured the magnetic filed H^z at the point $(x, y) = (-1, 0)$. In Figure 1 we compare the results obtained with both methods. The DG solution was obtained using third-order polynomials and a local refined mesh formed by 920 triangles. This means that the number of unknowns is 9.2×10^3. The FD solution was obtained using a uniform mesh with $h = 1.25 \times 10^{-2}$ implying 4.0×10^4 unknowns. Our experiments show that the DG method is more accurate and efficient. The error, in the Euclidean norm, is 1.5827×10^{-1} for the DG method and 3.6668×10^{-1} for the FD method.

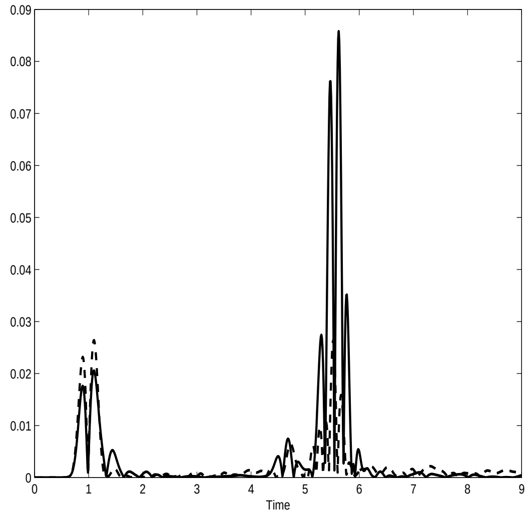


Figure 1: Error curves for the DG solution (dash line) and FD solution (solid line).

Acknowledgements

This work has been partially supported by 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 a Tecnologia under the project PEst-C/MAT/UI0324/2011, and by FCT under the project PTDC/SAU-ENB/119132/2010) and COMPETE (QREN -FCOMP-01-0124-FEDER-021014) program.

References

[1] R. Bernardes, T. Santos and J. Cunha-Vaz, *Evaluation of Blood-Retinal Barrier Function from Fourier Domain High-Definition Optical Coherence Tomography*, in IFMBE Proceedings 25/XI, 2009, 316–319.

[2] R. Bernardes, T. Santos, P. Serranho, C. Lobo, and J. Cunha-Vaz, *Noninvasive evaluation of retinal leakage using OCT*, Ophtalmologica **226(2)** (2011) 29–36.

[3] B. Bouma and G. Tearney, *Handbook of optical coherence tomography*, Marcel Dekker, New York, 2002.

[4] S. Descombes, S. Lanteri and L. Moya, *Locally implicit time integration strategies in a discontinuous Galerkin method for Maxwell's equations*, J. Sci. Comput. **56** (2013) 190–218.

[5] J. Hesthaven and T. Warburton, *Nodal High-Order Methods on Unstructured Grids: I. Time-Domain Solution of Maxwell's Equations*, J. Comput. Phys. **181** (2002) 186–221.

[6] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*, Springer Verlag, New York, 2008.

[7] L. Junqueira and J. Carneiro, *Basic Histology: Text & Atlas (Junqueira's Basic Histology)*, McGraw-Hill Medical, 2005.

[8] M. König, K. Busch and J. Niegemann, *The Discontinuous Galerkin Time-Domain method for Maxwell's equations with anisotropic materials*, Phot. Nano. Fund. Appl. **8** (2010) 303–309.

[9] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer Verlag, Berlim, 1969.

[10] J. Niegemann, R. Diehl and K. Busch, *Efficient low-storage Runge-Kutta schemes with optimized stability regions*, J. Comput. Phys. **231** (2012) 364–372.

[11] W. Reed, T. Hill, *Triangular mesh methods for the neutron transport equation*, Los Alamos Scientific Laboratory. LA-UR-73-479 (1973).

[12] R. N. Rieben, G. H. Rodrigue and D. A. White, *A high order mixed vector finite element method for solving the time dependent Maxwell equations on unstructured grids*, J. Comput. Phys. **204** (2005) 490–519.

[13] Z. S. Sacks, D. M. Kingsland, R. Lee and J.-F. Lee, *A perfectly matched anisotropic absorber for use as an absorbing boundary condition*, IEEE Trans. Antennas. Propag. **43** (1995) 1460–1463.

[14] J. M. Schmitt, A. Knüttel, and R. F. Bonner, *Measurement of optical properties of biological tissues by low-coherence reflectometry*, Appl. Opt. **32** (1993) 6032–6042.

[15] P. Serranho, M. Morgado and R. Bernardes *Optical Coherence Tomography: a concept review* In: Optical Coherence Tomography: A Clinical and Technical Update. R. Bernardes & J. Cunha-Vaz Eds., Springer-Verlag, 2012, 139–156

[16] A. Taflove and S. C. Hagness, *Computational Electrodynamic: The Finite-Difference Time-Domain Method (2nd ed.)*, Artech House, Norwood, MA, 2000.

[17] K. S. Yee, *Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media*, IEEE Trans. Antennas. Propag. **14** (1966) 302–307.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Parallel extrapolated algorithms for computing PageRank

Josep Arnal[1], Héctor Migallón[2], Violeta Migallón[1], Juan A. Palomino[1] and José Penadés[1]

[1] *Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Alicante, 03690 San Vicente del Raspeig, Alicante*

[2] *Departamento de Física y Arquitectura de Computadores, Universidad Miguel Hernández, 03202 Elche, Alicante*

emails: arnal@dccia.ua.es, hmigallon@umh.es, violeta@dccia.ua.es, japb4@alu.ua.es, jpenades@dccia.ua.es

Abstract

The PageRank algorithm for determining the importance of Web pages has become a central technique in Web search. This algorithm uses the Power method to compute successive iterates that converge to the principal eigenvector of the Markov chain representing the Web link graph. In this paper parallel Relaxed and Extrapolated algorithms based on the Power method are presented. Different parallel implementations of the Power method and the proposed variants are analyzed using different data distribution strategies. The reported experiments show the behavior of the designed algorithms for realistic test data using an hybrid OpenMP/MPI approach in order to exploit the benefits of shared memory inside the nodes of current SMP supercomputers.

Key words: PageRank, parallel algorithms, Power method, Relaxed and Extrapolated methods

1 Introduction

One of the most difficult problems in Web search is the ranking of the results recalled in response to a user query. Since contemporary Web crawls discover billions of pages, broad topic queries may result in recalling hundreds of thousand of pages containing the query terms. Only a few dozens of the recalled pages are actually presented to the user. Moreover, these results are presented in order of relevance. A variety of ranking features are used by Internet search engines to come up with a good order. Some of the query-independent features are based on page content. Some utilize the hyperlink structure of the Web. The PageRank algorithm [11] is one of those that introduces a content neutral-ranking function over Web pages. This ranking is applied to the set of pages returned by the Google search engine in response to posting a search query. PageRank is based in part on two simple common sense concepts: A page is important if many important pages include links to it and a page containing many links has reduced impact on the importance of the pages it links to. This model has been used by Google as part of its search engine technology. PageRank is essentially the stationary distribution vector of a Markov chain whose transition matrix is a convex combination of the Web link graph and a certain rank 1 matrix. A key parameter in the model is the damping factor, a scalar that determines the weight given to the Web link graph in the model. Due to the great size and sparsity of the matrix, methods based on decomposition are considered infeasible; instead, iterative methods are used, where the computation is dominated by matrix-vector products; see e.g., [1]. Traditionally, PageRank has been computed using the Power method. Many methods to accelerate the Power method have been developed such that extrapolation methods, block-structure methods or adaptive methods; see e.g., [6], [7], [8], [14] and the references cited therein. In recent years opportunities for parallel execution have broadened their scope; see e.g., [3], [4], [12]. In this paper different parallel algorithms based on the Power method for computing PageRank are proposed and analyzed. Concretely, we present parallel Relaxed and Extrapolated algorithms based on the Power method that accelerate its convergence. The remainder of the paper is structured as follows. In Section 2 we provide a brief description of the PageRank problem and we introduce the Power method and some acceleration methods based on the extrapolation technique. In Section 3, parallel algorithms of the Power method using relaxation and/or extrapolation are introduced. The numerical experiments performed in Section 4 show the behavior of these algorithms using different data distribution strategies on both shared and distributed memory multicore architectures. Finally, in Section 5 we give some conclusions.

2 Computing the PageRank

PageRank [11] is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. The PageRank problem can be seen as a matrix problem. Let $G = [g_{ij}]_{i,j=1}^{n}$ be a Web graph adjacency matrix with elements $g_{ij} = 1$ when there is a link from page j to page i, with $i \neq j$, and zero otherwise. Here n is the number of Web pages. From this matrix we can construct a transition matrix $P = [p_{ij}]_{i,j=1}^{n}$ as follows: $p_{ij} = \frac{g_{ij}}{c_j}$ if $c_j \neq 0$ and 0 otherwise, where $c_j = \sum_{i=1}^{n} g_{ij}$, $1 \leq j \leq n$, represents the number of out-links from a page j. For pages with a nonzero number of out-links, i.e., $c_j \neq 0$ for all j, $1 \leq j \leq n$, the matrix P is column stochastic. Thus each element of this matrix has values between 0 and 1, and the sum of the components of each column is 1. In this case the PageRank vector can be obtained by solving $Px = x$. Since we are interested in a probability distribution, the sum of the components of x is assumed to be one. Algorithm 1 shows the original Power method [13] for the PageRank computation where $e = (1, 1, \dots, 1)^T$. Note that we use the L_1 norm $\|x\|_1 = \sum_{i=1}^{n} |x_i|$.

Algorithm 1 *(Power method)*

```
Initialization $x^0 = \frac{e}{n},\ k = 0$
Repeat
    $x^{k+1} = Px^k$
    $\delta = \|x^{k+1} - x^k\|_1$
    $k = k + 1$
Until $\delta < \epsilon$
```

When the matrix $P \geq 0$ is irreducible (i.e., its graph is strongly connected) and stochastic, its largest eigenvalue in magnitude is $\lambda_{max} = 1$. Thus, Algorithm 1 converges to the eigenvector corresponding to $\lambda_{max} = 1$, and when normalized, it is the stationary probability distribution over pages under a random walk on the Web. However, the Web contains many pages without out-links, called dangling nodes. Dangling pages present a problem for the mathematical PageRank formulation because in this case the matrix P is non-stochastic and then Algorithm 1 can not be used. Moreover, the matrix irreducibility is not satisfied for a Web graph. In order to overcome this difficulty, Page and Brin [11] change the transition matrix P to a column stochastic matrix $\bar{P} = \alpha(P + vd^T) + (1 - \alpha)ve^T$, where $d \in \Re^n$ is the dangling page indicator defined by $d_i = 1$ if and only if $c_i = 0$ and the vector $v \in \Re^n$ is some probability distribution over pages. This model means that the random surfer jumps from a dangling page according to a distribution v. For this reason v is called a teleportation distribution. Originally uniform teleportation $v = \frac{e}{n}$ was used. Consequently, v is also known as a personalization vector. Then, setting α such that $0 < \alpha < 1$ (as a good compromise Google uses $\alpha = 0.85$), the matrix $\bar{P}$ is column stochastic and irreducible. On the other hand, since $\bar{P}$ is stochastic, then $\bar{P}$ preserves the L_1 norm, that is, $\|\bar{P}x\|_1 = \|x\|_1$ and therefore we can reformulate Algorithm 1 using the matrix $\bar{P}$ to solve the PageRank vector. That is, Algorithm 1 is utilized in order to solve the stationary distribution of the ergodic Markov chain defined by $\bar{P}$, $\bar{P}x = x$, obtaining the corresponding algorithm:

Algorithm 2 *(Power method for solving $\bar{P}x = x$)*

```
Initialization $x^0 = \frac{e}{n},\ k = 0$
Repeat
    $x^{k+1} = \alpha P x^k$
```

```
        $\gamma = \|x^k\|_1 - \|x^{k+1}\|_1$
        $x^{k+1} = x^{k+1} + \gamma v$
        $\delta = \|x^{k+1} - x^k\|_1$
        $k = k + 1$
    Until $\delta < \epsilon$
```

Note that although the matrix $\bar{P}$ is dense, Algorithm 2 has been designed such that it is not necessary to construct explicitly the matrix $\bar{P}$.

On the other hand, the extrapolation algorithms [5] accelerate the convergence of PageRank by using successive iterates of the Power method to estimate the nonprincipal eigenvectors of the hyperlink matrix, and periodically subtracting these estimates from the current iterate of the Power method. Concretely, the Extrapolation Power methods treated here exploit the knowledge of eigenvalues of the hyperlink matrix. Moreover, the extrapolation needs to be applied only once; see e.g., [5].

Algorithm 3 *(Extrapolation Power method)*

```
    Initialization $x^0 = \frac{e}{n},\ k = 0$
    Repeat
        $x^{k+1} = \alpha P x^k$
        $\gamma = \|x^k\|_1 - \|x^{k+1}\|_1$
        $x^{k+1} = x^{k+1} + \gamma v$
        If $k + 1 == r + 2$
            $x^{k+1} = \frac{x^{k+1} - \alpha^r x^{k+1-r}}{1-\alpha^r}$
        $\delta = \|x^{k+1} - x^k\|_1$
        $k = k + 1$
    Until $\delta < \epsilon$
```

3 Parallel Algorithms

In order to design the parallel algorithms, we consider that P is partitioned into p row blocks. Each block P_i, $1 \leq i \leq p$, is a matrix of order $n_i \times n$, with $\sum_{i=1}^{p} n_i = n$. Analogously, we consider the vectors x^k and v partitioned according to the block structure of P. Obviously, the Power method for solving $\bar{P}x = x$ can be executed in parallel. In this case each process actualizes a block of the vector x^{k+1} and a synchronization of all processes are performed at each iteration in order to construct the global iterate vector x^{k+1}. Due to this synchronization, we can use the formulation of Algorithm 2 because the property of preserving the L_1 norm remains valid.

Algorithm 4 *(Parallel Power method)*

```
Initialization $x^0 = \frac{e}{n},\ k = 0$
Repeat
  In process $i,\ i = 1, 2, \ldots, p$
      $x_i^{k+1} = \alpha P_i x^k$
      $\gamma = \|x^k\|_1 - \|x^{k+1}\|_1$
      $x_i^{k+1} = x_i^{k+1} + \gamma v_i$                                   (1)
  $x^{k+1} = [x_1^{k+1}, \ldots, x_p^{k+1}]$
  $\delta = \|x^{k+1} - x^k\|_1$
  $k = k + 1$
Until $\delta < \epsilon$
```

A relaxation parameter $\beta > 0$ can be introduced and replace the computation of x_i^{k+1} in (1) with the equation $x_i^{k+1} = \beta(x_i^{k+1} + \gamma v_i) + (1-\beta)x_i^k$. Clearly, with $\beta = 1$ equation (1) is recovered. In the case of $\beta \neq 1$ we have a Relaxed Power method. In order to accelerate the convergence of Algorithm 4 we also propose a parallel Relaxed Extrapolated algorithm based on Algorithm 3 as follows. In Algorithm 5 the relaxation is applied only after the extrapolation is performed.

Algorithm 5 *(Parallel Relaxed Extrapolated Power method)*

```
Initialization $x^0 = \frac{e}{n},\ k = 0$
Repeat
  In process $i,\ i = 1, 2, \ldots, p$
      $x_i^{k+1} = \alpha P_i x^k$
      $\gamma = \|x^k\|_1 - \|x^{k+1}\|_1$
      $x_i^{k+1} = x_i^{k+1} + \gamma v_i$
      If $k+1 == r+2$ then $x_i^{k+1} = \frac{x_i^{k+1} - \alpha^r x_i^{k+1-r}}{1-\alpha^r}$
      If $k+1 > r+2$ then $x_i^{k+1} = \beta x_i^{k+1} + (1-\beta)x_i^k$             (2)
  $x^{k+1} = [x_1^{k+1}, \ldots, x_p^{k+1}]$
  $\delta = \|x^{k+1} - x^k\|_1$
  $k = k + 1$
Until $\delta < \epsilon$
```

Note that the computation of $\|x^{k+1}\|_1$ and δ in Algorithms 4 and 5 is also performed in parallel in such a way that each process i computes the portions $\|x_i^{k+1}\|_1$ and $\|x_i^{k+1} - x_i^k\|_1$ followed by a reduction of these values. For the sake of simplicity, we have omitted these computations from the formulation of Algorithms 4 and 5.

4 Numerical setup

In order to illustrate the behavior of these methods, we have implemented the algorithms described here on an HPC cluster of 26 nodes HP Proliant SL390s G7 connected through a network of low-latency QDR Infiniband-based. Each node consists of two Intel XEON X5660 hexacore at up to 2.8 GHz and 12MB cache per processor, with 48 GB of RAM. The parallel environment has been managed using both MPI (Message Passing Interface) and OpenMP. That is, an hybrid MPI/OpenMP implementation has been designed by combining various OpenMP threads for each MPI process. That is, MPI is used for data distribution, and OpenMP in order to perform the computation inside the cores of each node. Concretely, let p be the number of processes performed, $p = s * c$ indicates that s nodes of the parallel platform have been used and for each one of these nodes, c OpenMP threads have been considered. Therefore, we use a philosophy of distributed shared memory using $p = s \times c$ processes or threads. Particularly, if $s = 1$ the algorithms are executed in shared memory by using $p = c$ threads on a single node. Conversely, if $c = 1$, we are working on distributed memory using $p = s$ nodes. In order to test our parallel implementations of the algorithms treated here we have used two datasets of different sizes, available from http://law.dsi.unimi.it [9]. These transition matrices have been generated from a web-crawl [2]. Table 1 summarizes, for each graph, the number of nodes (matrix size), the number of arcs (nonzero elements of the matrix), the density (arcs/nodes), the percentage of dangling nodes and the needed memory in order to store the matrix. As the dimension of the link matrix grows, its relative sparseness increases as well. To compute PageRank for large domains there is no possible way to work with the matrix in its full format, the memory requirements would be too high. Therefore, a sparse matrix format is needed in order to store the matrices. Concretely, the Compressed Sparse Row (CSR) format was used, which is one of the most extensively used storage scheme for general sparse matrices, with minimal storage requirements. We represent the two vectors of indexes of the CSR format by integers without sign of 32 bits, while the values are represented by using floating comma of 32 bits; the iterate vectors are represented by means of double precision of 64 bits. Taking into account that, for each column of the matrix, all nonzero elements are equal to a fixed value, it is stored once in an ordered vector. This modified CSR format (CSR') [10] has involved a reduction of memory requirements of about $63 - 73\%$ respect to the original CSR format. Implementing the PageRank calculations in a parallel environment

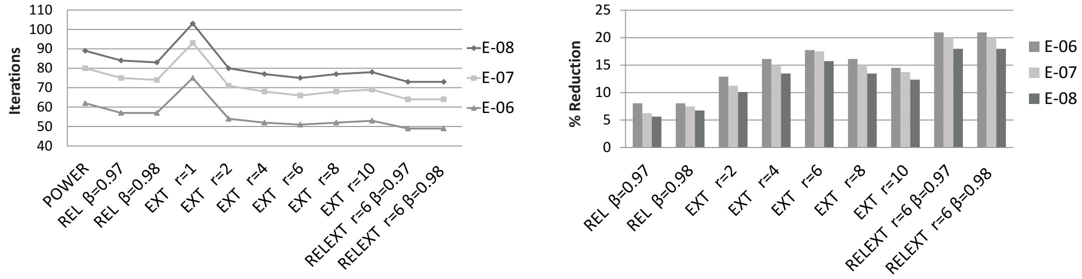
Graph	Nodes	Arcs	Dang. nodes	Density	Memory
it-2004	41,291,594	1,150,725,436	12.76%	27.87	4.75 GB
webbase-2001	118,142,155	1,019,903,190	23.41%	8.63	5.12 GB

Table 1: Graphs collection.

opens several possibilities in partitioning the data (i.e., how the data are divided among nodes) and load balancing (i.e., to ensure that all nodes perform similar amount of work). The most expensive operation performed in the calculation of the PageRank values is a matrix-vector multiplication. This is a perfectly parallel operation with several possible methods for partitioning both the matrix and the vector. We have considered two different methods for partitioning the link matrix among nodes. The first method we have chosen is a row-wise distribution (row-wise partitioning) where each node gets the same amount of rows. However, the number of nonzero elements per row of the link matrix used to calculate PageRank can differ immensely. In order to balance the calculations we consider a second matrix distribution strategy where each node has to handle the same amount of nonzero elements (nonzero elements partitioning).

4.1 Numerical results

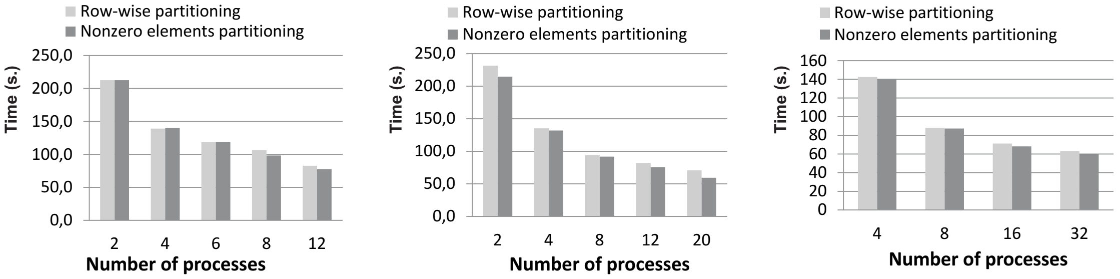
Using the datasets of Table 1 we illustrate the performance of the parallel algorithms proposed in this work for computing PageRank. Of the algorithms we have discussed here for accelerating the convergence of PageRank, the combining of relaxation and extrapolation performs the best empirically. Figure 1 compares the convergence rates for the Power method and the Relaxed (REL) and/or Extrapolated (EXT) methods setting a global convergence scheme and varying the stopping criterion for the matrix webbase-2001. As it can be seen in this figure, the proposed Relaxed Extrapolated (RELEXT) methods reduce the number of iterations needed to reach residuals of 10^{-8}, 10^{-7} and 10^{-6} by 18%, 20% and 21%, respectively. The stopping criterion used in the rest of the paper has been chosen such that $\epsilon = 10^{-8}$. For this sttoping criterion the Relaxed Extrapolated methods reduce the number of iterations respect to the Power method by 16% for the matrix it-2004. Note that the Simple Extrapolation ($r = 1$) is not effective (see Figure 1(a)) and slows down the convergence of the Power method. It is due to the fact that the Simple Extrapolation assumes that α is the only eigenvalue of modulus α and this is inaccurate; see e.g., [5]. By analyzing the performance of the two strategies of data distribution used, we have obtained that, generally, the nonzero elements partitioning is the best distribution strategy for all the algorithms treated here, more specially as the number of processes increases. Figure 2 illustrates this fact for the Power method for different number of processes using shared memory, distributed memory, and distributed shared memory. Our experience indicates that, for our datasets, a good choice of the values of r in the Extrapolated Power method (EXT) is $r = 6$, while good choices of β in the Relaxed Power method are between 0.97-0.98. Combining the relaxation and extrapolation such as is indicated in Algorithm 5, the best times have been obtained for these values of r and β. In order to deal with larger problems and to use all the available memory in a distributed system, the best strategy of parallelization need to use at the same time the benefits of shared and distributed memory multiprocessors. From the performed experiments it follows that usually the best parallel



(a) Number of iterations.

(b) Reduction respect to Power method (%).

Figure 1: Convergence rates for computing PageRank, webbase-2001.



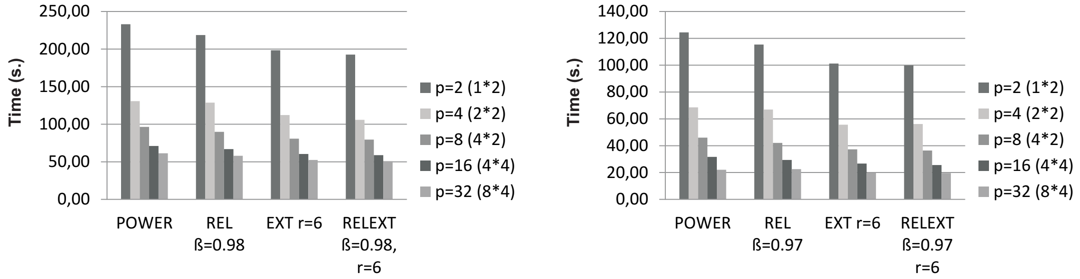
(a) Shared memory $p = 1 * c$.

(b) Distributed memory $p = s * 1$.

(c) Distributed Shared memory $p = s * 4$.

Figure 2: Row-wise versus nonzero-elements distribution. Power method, webbase-2001.

results are obtained using 1, 2 or 4 threads in each node. Moreover, it seems inappropriate to use 8 or more cores on a single node. Figure 3 shows the time that the described parallel algorithms take for computing PageRank varying the number of processes for the two datasets. The parallel Relaxed Extrapolated algorithms accelerate the convergence significantly saving $18 - 20\%$ in the time needed by the parallel Power method to reach a residual of 10^{-8} for the webbase-2001 and it-2004 matrices. On the other hand, the results show that a considerable speed-up is achieved. Concretely, the global efficiencies achieved respect to the sequential Power method were about $100 - 110\%$ for $p = 2$, $90 - 98\%$ for $p = 4$, $63 - 75\%$ for $p = 8$ and $45 - 55\%$ for $p = 16$, depending on the density of the matrices.



(a) webbase-2001.

(b) it-2004.

Figure 3: Parallel Relaxed and/or Extrapolated methods.

5 Conclusions

In this work we have made an analysis of parallel algorithms based on the Power method and the use of relaxation and/or extrapolation techniques for accelerating the computation of PageRank. Two strategies of data distribution have been used: row-wise partitioning and nonzero elements partitioning. An hybrid implementation has been designed by combining various OpenMP threads for each MPI process. The results show that the proposed parallel Relaxed Extrapolated algorithms can speed up convergence time significantly respect to the parallel Power algorithm.

Acknowledgements

This research was partially supported by the Spanish Ministry of Education, Culture and Sport under grant number TIN2011-26254.

References

[1] P. Berkhin, *A Survey on PageRank Computing*, Internet Mathematics **2(1)** (2005) 73–120.

[2] P. Boldi, B. Codenotti, M. Santini and S. Vigna, *Ubicrawler: A scalable fully distributed Web crawler*, Software: Practice & Experience **34** (2004) 711–726.

[3] D. Gleich, A. Gray, C. Greif and T. Lau, *An inner-outer iteration for computing PageRank*, SIAM journal of scientific computing **32(1)** (2010) 349–371.

[4] D. Gleich, L. Zhukov and P. Berkhin, *Fast Parallel PageRank: A linear system approach*, In The Fourteenth International World Wide Web Conference New York ACM Press, 2005.

[5] S. D. Kamvar, *Numerical Algorithms for Personalized Search in Self-organizing Information Networks*, Princeton University Press, 2010.

[6] S. D. Kamvar, T. H. Haveliwala and G. H. Golub, *Adaptive Methods for the Computation of PageRank*, Linear Algebra and its Applications **386** (2004) 51–65.

[7] S. D. Kamvar, T. H. Haveliwala, C. D. Manning and G. H. Golub, *Exploiting the Block Structure of the Web for Computing PageRank*, Stanford University Technical Report, SCCM-03-02, 2003.

[8] S. D. Kamvar, T. H. Haveliwala, C. D. Manning and G. H. Golub, *Extrapolation Methods for Accelerating PageRank Computations*, In Twelfth International World Wide Web Conference (2003) 261–270.

[9] Laboratoy for Web Algorithmics, law.dsi.unimi.it, 2002.

[10] H. Migallón, V. Migallón, J. A. Palomino and J. Penadés, *Parallelization Strategies for Computing PageRank*, In Proceedings of the Seventh International Conference on Engineering Computational Technology, Civil-Comp Press, Stirlingshire, UK, Paper 29, doi:10.4203/ccp.94.29, 2010.

[11] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank citation ranking: Bringing order to the Web*, Technical Report, Stanford Digital Library Technologies Project, 1999.

[12] A. Rungsawang and B. Manaskasemsak, *Parallel adaptive technique for computing PageRank*, In Proceedings of the 14th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP'06 (2006) 15–20.

[13] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford: Oxford University Press, 1998.

[14] G. Wu, Y. Wei, *An Arnoldi-Extrapolation algorithm for computing PageRank*, Journal of Computational and Applied Mathematics **234** (2010) 3196-3212.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Parallel evolutionary approaches for solving a planar leader-follower facility problem

A.G. Arrondo[1], J.L. Redondo[2], J. Fernández[1] and P.M. Ortigosa[3]

[1] *Department of Statistics and Operational Research, University of Murcia, Spain*

[2] *Department of Architecture and Technology, University of Granada, Spain*

[3] *Department of Informatics, ceiA3, University of Almería, Spain*

emails: agarrondo@um.es, jlredondo@ugr.es, josefdez@um.es, ortigosa@ual.es

Abstract

In the leader-follower facility problem considered in this work, the aim is to maximize the profit obtained by a chain (the leader) knowing that a competitor (the follower) will react by locating another single facility after the leader locates its own facility. The demand is elastic (it varies depending on the location of the facilities) and is supposed to be concentrated at n demand points. Several heuristic methods were proposed to cope with this hard-to-solve global optimization problem. Through a comprehensive computational study, it was shown that the evolutionary algorithm TLUEGO was the heuristic which provides the best solutions. Nevertheless, TLUEGO requires high computational effort, even to manage problems with small sizes. This is mainly due to the high cost at evaluating the leader's objective function, which requires the resolution of another hard-to-solve optimization problem, namely, the follower's problem. In this work, three parallel strategies of TLUEGO for solving this problem, i.e., a distributed memory programming model, a shared memory programming model and a hybrid of the two previous models, which can be executed in different architectures, are proposed.

Key words: Nonlinear bi-level programming problem, centroid (or Stackelberg) problem, evolutionary algorithm, high performance computing approaches
MSC 2000: AMS codes (optional)

1 Introduction

Competitive location deals with the problem of locating facilities to provide a service (or goods) to the customers (or consumers) of a given geographical area where other competing facilities offering the same service are already present (or will enter the market in the near future).

The scenario considered in this paper is that of a *duopoly*. A chain, the *leader*, wants to locate a new single facility in a given area of the plane, where there already exist m facilities offering the same goods or product. The first k of those m facilities belong to the chain ($0 \leq k < m$) and the other m-k to a competitor chain, the *follower*. The leader knows that the follower, as a reaction, will subsequently position a new facility too. The demand is supposed to be concentrated at n demand points, whose locations p_i are known. The location f_j and quality of the existing facilities are also known. The demand points split their buying power among the facilities proportionally to the attraction they feel for them. The attraction for a facility depends on both the location and the quality of the facility. In most competitive location literature it is assumed that the demand is deterministic, i.e., fixed regardless the conditions of the market. However, demand can vary depending on prices, distances to the facilities, etc. Taking variable demand into consideration, as we do here, increases the complexity of the problem and, therefore, the computational effort needed to solve it, but it makes some models more realistic. The aim is to maximize the profit obtained by the leader following the follower's entry. These types of bilevel programming problems were introduced by Hakimi, who introduced the terms *medianoid* for the follower problem, and *centroid* for the leader problem [2]. The interested reader is referred to [5] for an in-detailed description of the model.

For handling the centroid problem, several heuristics were proposed in [5]. The computational studies showed that the evolutionary algorithm TLUEGO (*Two-Level Universal Evolutionary Global Optimizer*) was more robust than the other strategies. However, the computational time employed by TLUEGO for solving small size problems was very high. This clearly suggests that a parallelization of the algorithm is needed, especially if real problems, with more demand points, are to be solved.

It is important to mention that to solve a single centroid problem, many medianoid problems have to be solved, since the evaluation of the leader's objective function at a given point requires the resolution of the corresponding medianoid problem. Recently, in [4], the medianoid problem considered in this work, has been studied and solved using the evolutionary algorithm UEGO (*Universal Evolutionary Global Optimizer*), initially described in [3]. The computational studies showed that the heuristic algorithm UEGO was a good alternative to deal with the medianoid problem, and hence it will be considered in this work. Nevertheless, solving a medianoid problem is not a negligible task; on the contrary, the medianoid problem is a hard-to-solve global optimization problem (as most competitive location problems are).

Traditionally, there have existed two parallel programming models: (i) shared memory programming, where the whole memory is directly accessible to all the processes with an intent to provide communication among themselves, and (ii) distributed programming, which is based on the message-passing mechanisms as a method to interchange information among processes. However, nowadays, computer systems are composed by several nodes with diverse processors which share memory; these nodes communicate to each other by using an interconnection network. This kind of architecture is called multicomputer. At this level, it may be possible to parallelize an application by using a hybrid programming model which combines both shared memory programming model and distributed programming model.

In this work, we propose three parallel approaches of TLUEGO, which make use of three parallel programming paradigms. The aim is to compare parallel versions and analyse theirs efficiency and effectiveness. The rest of the paper is structured as follows: In Section 2, the sequential TLUEGO is introduced and the main details of TLUEGO which can affect for its parallelization are described. Three parallel versions of TLUEGO are depicted in Section 3, and some computational studies are presented in Section 4. The paper ends with some conclusions in Section 5.

2 The sequential TLUEGO algorithm

TLUEGO is a multimodal evolutionary optimization algorithm based on subpopulations (denominated species). Initially, a random single species is generated considering the whole search space. Then, during the optimization process, a list of species is kept by TLUEGO. In fact, TLUEGO can be described as a method for managing this list (i.e. creating, deleting and optimizing species). As the algorithm evolves and applies genetic operators, new species can be created. TLUEGO performs a local optimizer operation on each species in the list, which enables the search to focus on the promising regions of the space. At each iteration of the algorithm, a selection procedure is carried out in order to keep only the best species in the list.

It is important to mention that there is no relationship among species, which means that a single species can create new candidate solutions and evolve without participation of the remaining ones. Therefore, there exists an intrinsic parallelism that can be exploited by dividing the species list into the processors.

Notice also that computational load of TLUEGO is concentrated in the evaluation of the objective function after the creation of new candidate solutions and in the local optimization procedure. Therefore, both methods are suitable for being run in parallel. The selection procedure is a very fast method which requires the knowledge of the whole species list. So, it is necessary that a single processor performs it in order to proceed as the sequential algorithm.

3 The parallel TLUEGO approaches

In this section, three parallel strategies of TLUEGO are briefly described.

The first parallel approach of TLUEGO is based on message-passing mechanisms (MPI, [6]) and implements a *master-slave technique*. Master-slave is a communication model where one processor (the master) has unidirectional control over one or more processors (the slaves). This technique is called "global parallel model" too, since the management of the population, is global (i.e. all the individuals in the population are considered when the selection procedure is carried out).

The second parallel approach considers a shared memory programming model. In this case, OpenMP has been selected [7] and the parallel strategy can be considered a *pseudo* master-slave technique, since there does no exist a master processor which distributes the species in the list; OpenMP includes mechanisms to manage the whole species list by all the processors in parallel, although a single processor is in charge of selecting the best species in the list.

The last parallel version of TLUEGO combines both message-passing and shared memory programing models. The result is a hybrid parallel model which exploits parallelism beyond a single level. A *coarse-grain model* is considered in an upper level, where the species list is distributed, and the pseudo master-slave strategy described above is used to work with the corresponding species sublists in a lower level. Notice that MPI is need to communicate species in the upper level and OpenMP is used for creating, deleting and optimizing species of the sublists.

4 Computational studies

All the computational study are carried out in the supercomputer Ben Arabi of the Supercomputing Center of Murcia, Spain. In particular, the executions run in Arabi, which is a Blade Cluster with 816 cores, organized in 32 nodes with 16GB of memory each, and 70 nodes with 8GB each (102 nodes altogether). Each node has 8 cores, divided into 2 Intel Xeon Quad Core (E5450) to 3.0 GHz. The algorithms have been implemented in C++.

In order to study the performance of the parallel algorithms, different types of problems, varying the number n of demand points, the number m of existing facilities and the number of k of those facilities belonging to the leader chain were generated. Moreover, several configurations varying the number of nodes, N, of Arabi and the number of cores used inside each node, N_c, have been considered.

To measure how well-utilized the processors are in solving the problem when the parallel implementations use P processors, the *efficiency* metric has been used, $Eff(P) = \frac{T(1)}{P \cdot T(P)}$. Notice that $T(1)$ is the time employed by the sequential TLUEGO, and the value of P is given by the number N of nodes multiplied by the number N_c of processors used in the nodes.

Our preliminary studies show that all parallel strategies have a good behavior in terms of effectiveness, since they always find the same solution as the sequential algorithm. Notice that, due to the stochastic nature of the algorithms, all the experiments have been run 10 times for each instance and average values have been computed.

Concerning the efficiency analysis, both the MPI pure master-slave strategy and the OpenMP pure pseudo master-salve technique reduce the computational time proportionally to the number of processors, up to 8 processors. However, the efficiency decreases when the number of processors P increases. This behaviour also appears in the hybrid parallel version, in which can be observed poor efficiency values as the number of nodes, N, increases.

5 Conclusion

In this study, a leader-follower problem with variable demand presented in literature (see [5]) is considered. In the problem, a chain (the leader) has to decide where to locate a new facility (and its quality) knowing that a competitor (the follower) will react by locating another facility. The demand is assumed to be variable, depending on the distances to and on the quality of the facilities. The model is a hard-to-solve global optimization problem.

Three parallel implementations of TLUEGO, devised to be executed in different architecture platforms, have been proposed. All of them exhibit good performance behaviors, obtaining either optimal or near-optimal efficiency for up to 8 processors. The results are promising and, in the near future, the scalability of these parallel approaches will be studied deeper.

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Science and Innovation (TIN2008-01117, ECO2011-24927), Junta de Andalucía (P10-TIC-6002), Program CEI from MICINN (PYR-2012-15 CEI BioTIC GENIL, CEB09-0010) and Fundación Séneca (The Agency of Science and Technology of the Region of Murcia, 00003/CS/10 and 15254/PI/10), in part financed by the European Regional Development Fund (ERDF).

References

[1] B. Chapman, G. Jost and R. van der Pas. Using OpenMP. Portable shared memory parallel programming. The MIT Press, Cambridge, Massachusetts, 2008.

[2] S.L. Hakimi. On locating new facilities in a competitive environment. *European Journal of Operational Research*, 12(1):29–35, 1983.

[3] M. Jelásity, P.M. Ortigosa, and I. García. UEGO, an abstract clustering technique for multimodal global optimization. *Journal of Heuristics*, 7(3),215–233, 2001.

[4] J.L. Redondo, J. Fernández, A.G. Arrondo, I. García, and P.M. Ortigosa. Deterministic or variable demand? Does it matter when locating a facility? Omega, 40(1): 9–20, 2012.

[5] J.L. Redondo, A.G. Arrondo, J. Fernández, I. García, and P.M. Ortigosa. A Two-Level Evolutionary Algorithm for Solving the Facility Location and Design ($\mathbf{1}|\mathbf{1}$)-centroid problem on the Plane with Variable Demand. Submitted

[6] M. Snir, S. Otto, S. Huss-Lederman, D. Walker and J. Dongarra. MPI: The Complete Reference. The MIT Press, Cambridge, Massachusetts. London, England, 1996.

[7] B. Chapman and G. Jost and R. van der Pas. Using OpenMP. Portable shared memory parallel programming. The MIT Press, Cambridge, Massachusetts. 2008.

*Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.*

Analysis of the Performance of the Google Earth Running in a Cluster Display Wall

Ismael Arroyo[1], Francesc Giné[1] and Concepció Roig[1]

[1] *Computer Science Department, University of Lleida*

emails: ismael.arroyo@udl.cat, {sisco, roig}@diei.udl.cat

Abstract

Nowadays, clusters of interconnected workstations have become a common solution for powering large composite displays, or "cluster display walls". Our paper is focused on analyzing a specific cluster display wall developed by Google Technology, named Liquid Galaxy, made up of homogeneous commodity hardware, running the well-known Google Earth application. With this in mind, we defined and tested different scenarios, representing the behavior of many kinds of user navigating around different places with variable densities of 3D-buildings. Our results show that the CPU, memory and local network are good enough to visualize the images properly, while depending on the user behavior, the external network constitutes the bottleneck of the system.
Key words: Cluster Display Wall, Liquid Galaxy, Google Earth.

1 Introduction

A current method for visualizing large-scale data sets consists of using a display wall based infrastructure, in which various screens are distributed in tiles connected to a single computer using multiple video outputs. The main problem that arises when using display walls is that the images are stretched, as resolution of a single screen is resized to fit all the screens together. To bypass this problem, a cluster-based infrastructure, in which each screen is served by a computer, is proposed. Some examples are CAVE [1] and GeoWall [2]. In this case, visualization of the images accross screens is carried out by using synchronization between computers to pass the data between them. This kind of system is capable of higher resolution and, consequently, it facilitates the analysis of the visualized data.

Nowadays, besides the fact that conventional PCs can be equipped with powerful consumer graphics cards with multiple outputs, the availability of software packages for clusters makes setting up cluster display walls affordable. This opens a wide range of new possibilites for using them in day-to-day scenarios in such different areas as entertainment, finance or education.

This paper focuses on the use of a specific cluster display wall hemisphere infrastructure developed by Google, named Liquid Galaxy [3]. This is composed of eight displays each connected to a computer node. It was conceived for executing Google Earth[4] and to provide an immersive geographical visualization. Taking into account the architecture of Liquid Galaxy and its possibilities, we built a new cluster display wall infrastructure using commodity hardware to identify performance issues for use in a wider range of applications.

We evaluated the main performance metrics (CPU, memory and networking) by running Google Earth in different scenarios. The results show that a cluster display wall composed of conventional desktop computers provides enough CPU and memory to execute Google Earth. However, networking constitutes an important issue, as we identify the user behavior as the determining factor in the bandwidth requirements. This analysis will permit our future work to be focused on identifying new applications fields and scenarios in which this new infrastructure can be used to visualize data and facilitate its understanding.

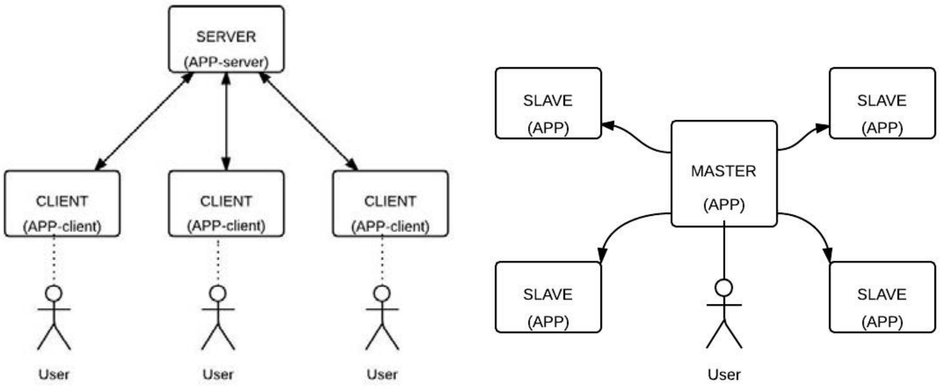
The paper is structured as following. Section 2 presents the architectures used in cluster display walls. Section 3 presents the case of study of Liquid Galaxy. In Section 4, we evaluate the performance parameters of a Liquid Galaxy infrastructure built with commodity hardware and also identify performance issues. Finally, in Section 5, we conclude the paper and discuss future directions.

2 Cluster Visualization Systems

A display wall is a renderization system with multiple screens connected to a single computer. Having one computer per screen is the main difference between a display wall and a cluster based display wall, because it has higher resolution by far. It allows a great amount of data to be visualized simultaneously in high definition, making it suitable for a wide range of applications where very high resolution is needed. Some current cluster display walls are: CAVE [1], GeoWall [2] and Liquid Galaxy [3]. Also, many works have focused on enabling collaboration between users through screen sharing in the same room, for example: SAGE [5], ViewDock TDW [6], Dynamo [7], Impromptu [8] and WeSpace [9]

Depending on the configuration of the system and the communication protocol to be applied, we can distinguish the following two cluster display wall architectures [10].

- **Client-server:** In the client-server model (Figure 1a), there are two sides of the application running on the system. The server is running the server-side application (APP-server), which is the main application, and which stores and refreshes the clients' status. Each client runs an instance of the client-side application (APP-client), which is connected to the server and modifies their status and/or some of the content of the APP-server. The user can choose one of the client nodes to interact with. This architecture can be used to execute both interactive and non-interactive applications. Interactive applications are those that allow a user to access a client node and modify its status. Consequently, it will send the changes to the server making the others clients fetch these new data and refreshing the stored status for that client node. OptiPlanet [11], Equalizer [12], Chromium [13] and CGLX [14] are examples of interactive applications to be executed in a cluster display wall environment. The non-interactive applications do not usually allow any changes by the user, although, in the case that the user could make some, these would not affect either the server or the other clients. An example of a non-interactive application is Video Streaming [15].



(a) Client-Server (b) Master-Slave

Figure 1: Cluster display wall architectures

- **Master-Slave:** In the Master-Slave architecture (Figure 1b), each node runs the same application (APP), so the master must manage the synchronization among slaves to ensure consistency. The master is the only node that accepts user input and every time that the master modifies its status, it sends the changes to the slaves. Applications for this kind of architecture are usually interactive, such as Google Earth.

In this paper, we focus on the Master-Slave architecture. We study the performance characteristics of this kind of system with the specific platform, Liquid Galaxy, which is implemented in a cluster display wall, developed by Google, to run Google Earth. In the next section, we present the characteristics of the Liquid Galaxy system.

3 Liquid Galaxy

Liquid Galaxy [3] is a cluster display wall originally built to run Google Earth [4] in order to create an immersive experience for the user. Liquid Galaxy lets you navigate around the globe with its 6-axis controller, allowing you instantly zoom in, out, and turn around with completely fluid motion. You can also search and navigate to specific locations on autopilot using a touch-screen interface. Figure 2 shows an example of the Google Liquid Galaxy installed in the Technological Park in the city of Lleida (Spain) [16].

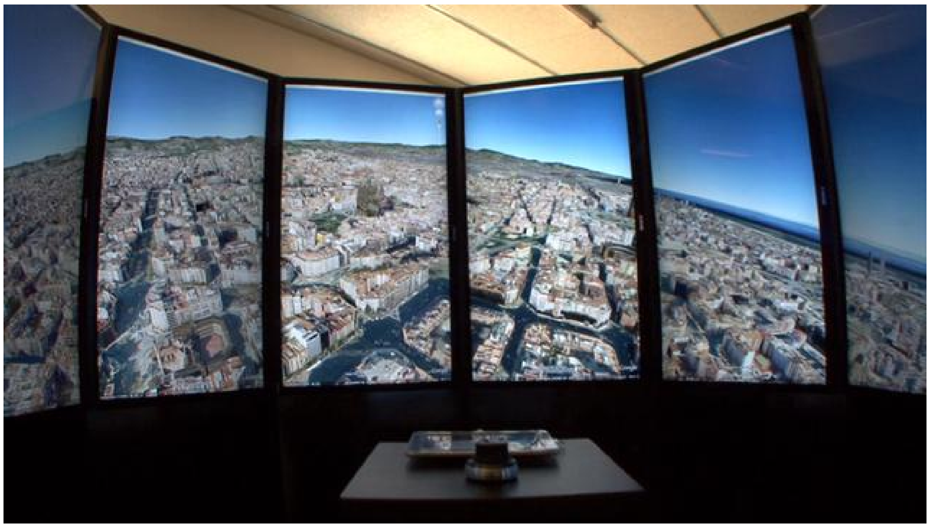


Figure 2: Lleida Liquid Galaxy

The connection schema of the Liquid Galaxy is depicted in Figure 3, which also indicates the different running steps.

We can see that each node runs the Google Earth application (GE) and the user only interacts with the master node by means of a 3D mouse. Each movement of the mouse makes the master node launch a synchronization protocol that consists of the following steps:

1. The master node captures the coordinates (Coords) of the position in Google Earth indicated by the user. These coordinates are codified in a UDP packet, named ViewSync, which contains the following information from the view in the application: latitude, longitude, altitude, heading, tilt, roll and planet name.

2. When the master's view is moved, it sends ViewSync packets to broadcast with the objective of sending it to all the nodes in the network.

3. Every slave node has a configuration file, which holds an offset from the original

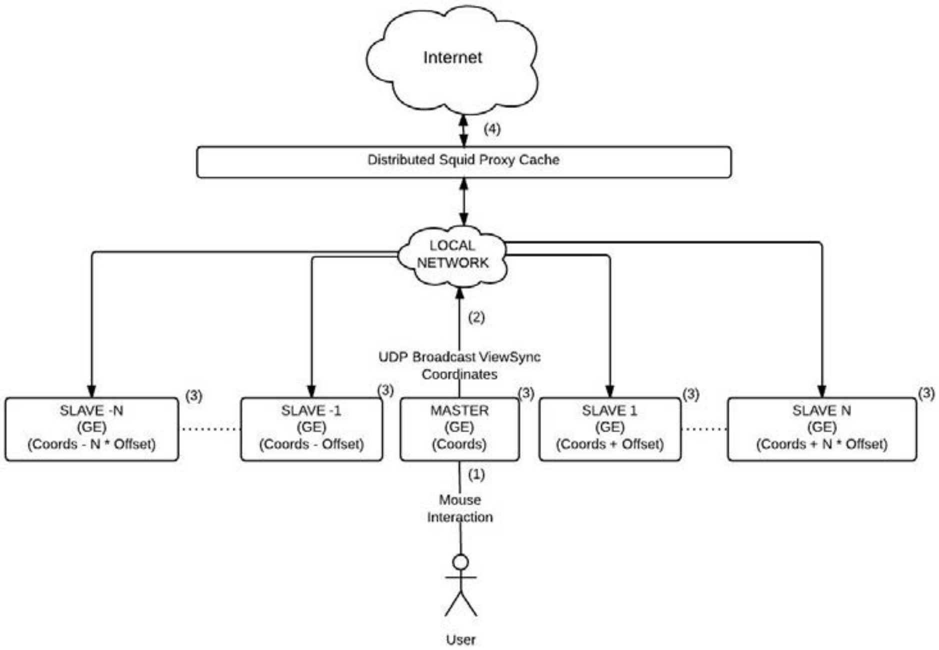

Internet
(4)
Distributed Squid Proxy Cache
LOCAL NETWORK
(2)
UDP Broadcast ViewSync Coordinates
SLAVE -N (GE) (Coords - N * Offset)
(3)
SLAVE -1 (GE) (Coords - Offset)
(3)
MASTER (GE) (Coords)
(3)
SLAVE 1 (GE) (Coords + Offset)
(3)
SLAVE N (GE) (Coords + N * Offset)
(3)
(1)
Mouse Interaction
User

Figure 3: Google Earth Architecture

master's view. When the slaves receive a ViewSync packet from the master, they automatically calculate and adjust their relative view by adding the local offset.

4. Every node accesses Internet with its own coordinates to download the required data (maps, imagery, 3D layer, etc.) independently from the other nodes.

As we have seen in previous steps, every node executes its own instance of Google Earth, so every node needs an Internet connection to download all the data. For this reason, a web-proxy distributed cache fits very well. Squid [17] is the official distributed cache and is also included in Liquid Galaxy. All nodes will share the same disk cache stored in each node's SSD. With this solution, the amount of data requests from Internet can be reduced. So, if the information is available in the distributed Squid cache, shared by all nodes as peers, it is taken independently by each node.

The Liquid Galaxy system presented in this section was built specifically to give service to run Google Earth. However, the immersive visualization environment that Liquid Galaxy provides opens this kind of system for use in a wide range of applications that can benefit from an immersive visualization environment. Some examples of applications that can be run in this system are WebGL with Aquarium [18], video streaming [15], videogames like Quake 3 Arena [19], etc.

Taking this possibility into account, we built a new experimental infrastructure that we named Liquid Galaxy Code. This new platform is composed of commodity hardware with the aim of enabling it to be built economically thus ensuring extensive use. An experimentation process was carried out on this platform to evaluate its viability. This process is reported in the next section.

4 Experimentation

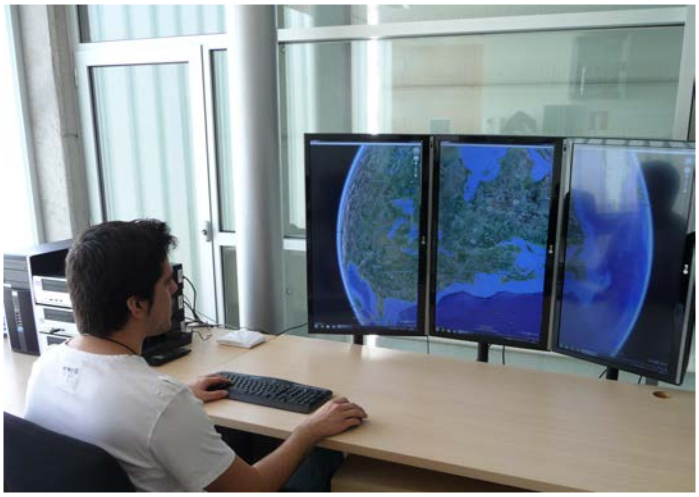
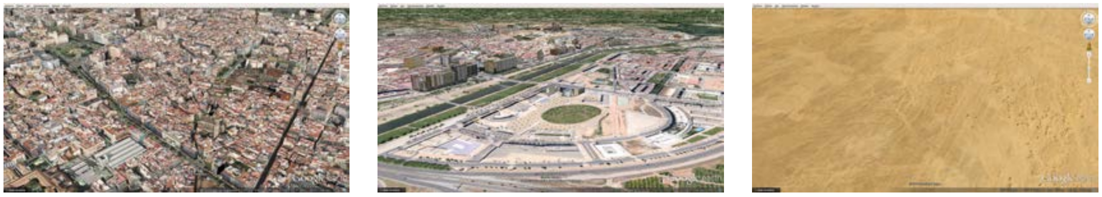


Figure 4: Liquid Galaxy Code

In this section, an experimentation process is carried out to measure and categorize the overhead produced by the use of Google Earth in the Liquid Galaxy Code infrastructure in order to show the resource requirements in usual displacements along the globe. In Figure 4, we can see the experimentation platform, which is a cluster display wall of 3 Intel Core i5 3GHz machines, with 2x4GB RAM 1600 MHz, SSD 128GB, NVidia GT620 with three 32" screens. One node acts as a master and the other two as slaves. This platform was built with commodity hardware in order to evaluate the viability of using general-purpose platforms to visualize high-resolution images in a cluster based way.

For the application under study, the requirements for computing resources vary with the environment to be visualized. Thus, we measured the three following environments, illustrated in Figure 5, with significant differences in the number of high-defined 3D buildings to be visualized:

- *City:* A high-density place made up by 3D buildings.



(a) City tour (b) Town tour (c) Desert tour

Figure 5: Images of the three tested environments

- *Town:* A low to medium-density place that combines zones with few 3D buildings and areas without them.
- *Desert:* A low-resolution without any 3D buildings to be loaded.

Every environment, or tour, moves throughout eight specific points of interest with a previously-established time interval between consecutive jumps. We distinguish the two following intervals:

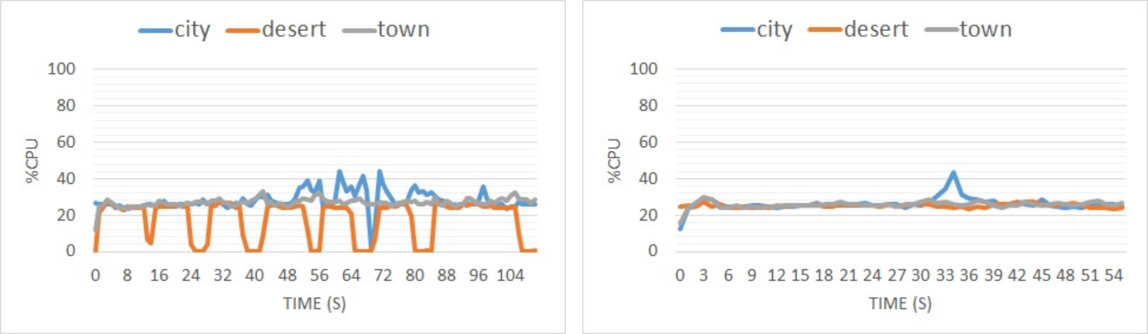
- Long time jumps: the time interval for moving between two consecutive places is 14 seconds.
- Short time jumps: the time interval has been set to 7 seconds.

Under the experimentation framework presented above, the key computing resources, CPU, memory and networking were monitored using the Top and Tshark (Command-line based WireShark) tools. The results are presented below.

4.1 Experimental Results

In our experimentation, both the master and slave nodes were monitored. The results obtained showed that there are no difference between the master node and the slaves. For this reason, we only present the results obtained from the master node.

Figure 6 shows the percentage of CPU usage for the three described environments and both timing jumps. In the short time jumps, there was no difference between the three environments, and the use of CPU averaged 25% for all of them. This is due to the fact that there was not enough time between jumps to load the needed data. On the other hand, for the long time jumps, there was a remarkable difference depending on the kind of environment. The big difference was in the desert test where there were peaks at each idle position. This was caused by the low-resolution imagery and the lack of 3D buildings. This made the time interval between jumps long enough to allow the images to be fully loaded. Referring to the other environments, city and town, similar behavior can be seen

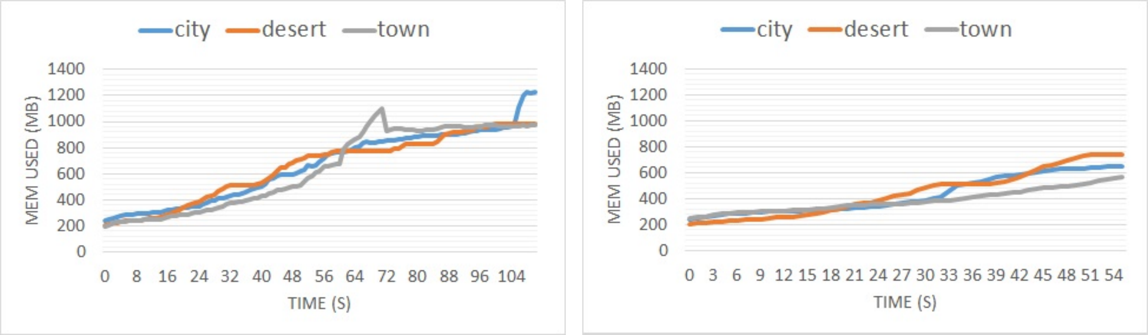


(a) Long time jumps

(b) Short time jumps

Figure 6: Percentage of CPU usage

in the CPU usage, because there are 3D buildings in both environments. However, CPU usage was slightly higher in the city due to the higher density of 3D buildings. According to these results, there is no need for a very high-performance CPU for the nodes, given that the average usage was below 30% in all cases. Nevertheless, the time between jumps is determinant for fully visualizing the images of the environment.



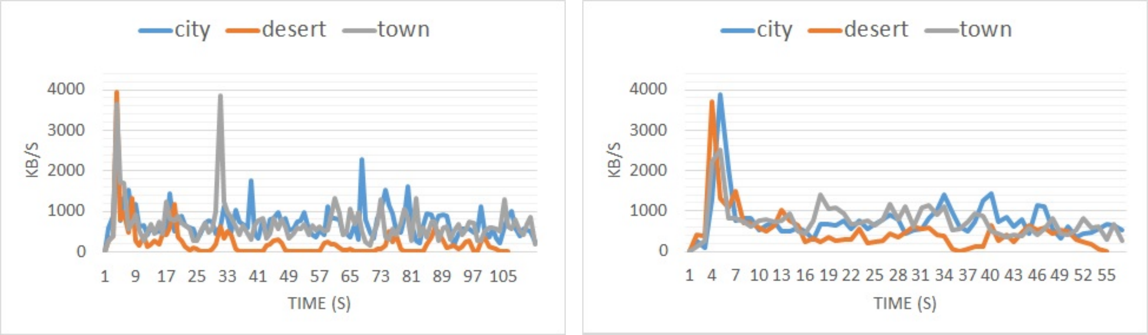
(a) Long time jumps

(b) Short time jumps

Figure 7: Memory usage

Another parameter studied was the memory usage throughout the test. Figure 7 shows this parameter under the same experimental framework described above. All tests proved that the RAM always increases moderately regardless of the environment and the timing jumps. Memory will always be reserved at the same rate until the limit established by the application is reached. This limit is always less than the total memory available.

A key performance parameter to analyze is the network usage because Google Earth has a heavy request of data from Internet and many synchronization packets per second. Figure



(a) Long time jumps

(b) Short time jumps

Figure 8: Networking.

8 presents the network graph measured in kilobytes/s. Despite having the Squid cache [17] activated, we can still perceive peaks whenever a request for data is made. This happens because Squid only has an average of 20% cache hits, leading the nodes to download new data when the next place to visit is far away. There is also a common starting peak in both timings and all the tours. This initial peak appears because the tour starts with a general view of the Earth and has to move to the first coordinate, having to load all the data until the first position is reached. We can also note that the desert tour has a lower data-request because it has no 3D buildings and low-resolution maps to download. Regarding the graph of short time jumps, the time between jumps is so close that the 3D layer has almost no time to load, making it unnecessary to request that heavy-load data. The synchronization between slaves is very abundant but it does not have very much weight, making it unnoticeable for the network. So, the local network does not need a high bandwidth.

5 Conclusions and Future Work

In this paper, we have presented a cluster display wall infrastructure, named Liquid Galaxy Code, which has been built with commodity hardware. We analyzed the performance of this system running Google Earth, so that the results may be indicative of the viability of using this kind of infrastructure in a wide range of everyday applications. These applications can benefit from an immersive visualization environment showing the viability of using the tested platform in a more extensive way.

The performance analysis have shown that CPU, memory and local network are enough to visualize the images. However, the bandwidth of the external network constitutes a bottleneck. This causes that the user navigation behaviour in the tested environments, is determinant in order to be able to load all the needed data. Thus, the time interval between consecutive displacements has to be adjusted according to the density of data of the environment to be visualized.

In the future works, we plan to indentify new scenarios and applications from different fields, such as education, finance, chemistry or biology, which can benefit from the presented platform built with off-the-shelf hardware. Additionally, we will study how to adjust and synchronize a cluster display wall composed of heterogeneous nodes.

Acknowledgements

We would like to thank to the company Ponent 2002 for facilitating the use of the Lleida Liquid Galaxy and specifically to Andreu Ibañez for his collaboration in understanding its operation.

References

[1] T. A. DeFanti, D. Acevedo, R. A. Ainsworth, M. D. Brown, S. Cutchin, G. Dawe, K.-U. Doerr, A. Johnson, C. Knox, R. Kooima, *et al.*, "The future of the CAVE," *Central European Journal of Engineering*, vol. 1, no. 1, pp. 16–37, 2011.

[2] A. Johnson, J. Leigh, P. Morin, and P. Van Keken, "Geowall: Stereoscopic visualization for geoscience research and education," *Computer Graphics and Applications, IEEE*, vol. 26, no. 6, pp. 10–14, 2006.

[3] Google, "Liquid Galaxy live demo at TED." Lecture Notes in Computer Science, Springer, 2010.

[4] Google, "Google earth," 2013. http://www.google.com/earth/index.html.

[5] L. Renambot, A. Rao, R. Singh, B. Jeong, N. Krishnaprasad, V. Vishwanath, V. Chandrasekhar, N. Schwarz, A. Spale, C. Zhang, *et al.*, "SAGE: the Scalable Adaptive Graphics Environment," in *Proceedings of WACE*, vol. 9, pp. 2004–09, Citeseer, 2004.

[6] C. D. Lau, M. J. Levesque, S. Chien, S. Date, and J. H. Haga, "ViewDock TDW: high-throughput visualization of virtual screening results," *Bioinformatics*, vol. 26, no. 15, pp. 1915–1917, 2010.

[7] S. Izadi, H. Brignull, T. Rodden, Y. Rogers, and M. Underwood, "Dynamo: a public interactive surface supporting the cooperative sharing and exchange of media," in *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pp. 159–168, ACM, 2003.

[8] J. T. Biehl, W. T. Baker, B. P. Bailey, D. S. Tan, K. M. Inkpen, and M. Czerwinski, "Impromptu: a new interaction framework for supporting collaboration in multiple display environments and its field evaluation for co-located software development," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 939–948, ACM, 2008.

[9] D. Wigdor, H. Jiang, C. Forlines, M. Borkin, and C. Shen, "WeSpace: the design development and deployment of a walk-up and share multi-surface visual collaboration system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1237–1246, ACM, 2009.

[10] C. N. Oliver G. Staadt, Justin Walker and B. Hamann, "A survey and performance analysis of software platforms for interactive cluster-based multi-screen rendering," *Eurographics Workshop on Virtual Environments*, 2003.

[11] L. Smarr, M. Brown, and C. de Laat, "Special section: OptIPlanet—the OptIPuter global collaboratory," *Future Generation Computer Systems*, vol. 25, no. 2, pp. 109–113, 2009.

[12] S. Eilemann, M. Makhinya, and R. Pajarola, "Equalizer: A scalable parallel rendering framework," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 3, pp. 436–452, 2009.

[13] G. Humphreys, M. Houston, R. Ng, R. Frank, S. Ahern, P. D. Kirchner, and J. T. Klosowski, "Chromium: a stream-processing framework for interactive rendering on clusters," in *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 693–702, ACM, 2002.

[14] K. Doerr and F. Kuester, "CGLX: a scalable, high-performance visualization framework for networked display environments," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 3, pp. 320–332, 2011.

[15] "Panoramic Video Streaming."
https://code.google.com/p/liquid-galaxy/wiki/PanoramicVideo.

[16] P. 2002, "G-liquid galaxy," 2012. http://www.g-liquidgalaxy.com/.

[17] T. Squid, "Squid web proxy cache," 2002.

[18] Greggman and H. Engines, "Aquarium WebGL."
http://webglsamples.googlecode.com/hg/aquarium/aquarium.html.

[19] OpenArena Team, "OpenArena (Quake 3 Arena)," 2005. www.openarena.ws.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

High-order iterative methods by using weight functions technique

Santiago Artidiello[1], Alicia Cordero[2], Juan R. Torregrosa[2] and Maria Vassileva[1]

[1] *Ciencias Básicas y Ambientales, Instituto Tecnológico de Santo Domingo (INTEC), Dominican Republic*

[2] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Spain*

emails: santiago.artidiello@intec.edu.do, acordero@mat.upv.es, jrtorre@mat.upv.es, maria.penkova@intec.edu.do

Abstract

In this work, two families of optimal iterative methods of order eight, for solving nonlinear equations, are presented. We design them by using the weight function technique, with functions of one, two and three variables. The study of dynamical and numerical behavior of some elements of these families complete this paper.

Key words: Nonlinear equation, Iterative schemes, optimal methods, weight functions technique, efficiency index
MSC 2000: 65H05, 37F10

1 Introduction

Many problems in science and engineering include the search of the solutions of an equation of nonlinear nature. The importance of this type of troubleshooting has showed the need to develop iterative methods for the resolution of these equations.

We consider the problem of finding a zero of a nonlinear function $f : I \subset \mathbb{R} \to \mathbb{R}$, that is, a solution $\xi \in I$ of

$$f(x) = 0, \tag{1}$$

restricted to real functions with a unique solution inside an open interval I. Currently there exist numerous iterative methods for solving the nonlinear equation (1). The majority of the computationally useful techniques to determine roots can be classified into one of the two classes: (a) iterative methods that require only functional evaluations of f, and (b) methods whose iterative formula require evaluations of the function and its derivatives. There are two known simple and effective methods that represent these classes: Steffensen' and Newton's methods, respectively, both with order of convergence two (see, for example [6]). The search of variants of these methods with an accelerated convergence and a reduced number of operations and functional evaluations has resulted in multistep methods. These schemes belong to the class of most powerful methods that overcome the limitations of the methods of a point with respect to the order of convergence and computational efficiency. This type of iterative methods were extensively studied in Traub's book, [11] and more recently, in [8].

Newton's method is among the most used procedures. Many authors have studied and proposed various multistep methods that modify the mentioned scheme to solve nonlinear equations with a high-order of convergence (see [4, 3, 2]). However, these must use derivatives, which is a serious disadvantage. Sometimes, the applications of iterative algorithms that depend on derivatives are restricted in engineering and science. For these cases, different authors have developed derivative-free iterative methods in numerous papers, that only need the additional evaluations of the function (see [12, 9, 5, 10]).

In this paper, following the ideas presented in [7] and [1], we present two families of iterative methods belonging to classes (a) and (b), respectively, both of eighth-order of convergence.

2 Design of the families

Let us remember that the iterative expression of Steffensen's scheme is

$$x_{k+1} = x_k - \frac{f(x_k)^2}{f(x_k + f(x_k)) - f(x_k)} \tag{2}$$

and Newton's one is

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{3}$$

If we compose any of these methods with itself, schemes of order 4 with 4 functional evaluations are obtained in both cases. That makes them non-optimal, according to the conjecture of Kung-Traub [13].

In double Newton scheme, let us consider a frozen derivative in the second step, obtaining a non-optimal method of order 3 (see [11]), whose iterative expression is:

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= y_k - \frac{f(y_k)}{f'(x_k)}. \end{aligned} \tag{4}$$

The authors in [1], using the technique of weight functions, got to increase to four the order of convergence of (4) without adding new functional evaluations. The iterative scheme of the obtained method is

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= y_k - H(u_k)\frac{f(y_k)}{f'(x_k)}, \end{aligned} \tag{5}$$

where $H(u_k)$ represents a real-valued function and $u_k = \dfrac{f(y_k)}{b_1 f(x_k) + b_2 f(y_k)}$, being b_1 and b_2 real parameters.

Now, let us consider the double Steffensen scheme, holding the expression of the denominator in the second step

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[z_k, x_k]}, \\ x_{k+1} &= y_k - \frac{f(y_k)}{f[z_k, x_k]}, \end{aligned} \tag{6}$$

where $z_k = x_k + f(x_k)$ and $f[z_k, x_k] = \dfrac{f(z_k) - f(x_k)}{z_k - x_k}$. By using the technique of weight functions, Petković et al. in [7] increase to four the order of convergence of method (6) without adding new functional evaluations. The iterative scheme of the resulting method is

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[z_k, x_k]}, \\ x_{k+1} &= y_k - H(u_k, v_k)\frac{f(y_k)}{f[z_k, x_k]}, \end{aligned} \tag{7}$$

where $z_k = x_k + \beta f(x_k)$, $u_k = \dfrac{f(y_k)}{f(x_k)}$ and $v_k = \dfrac{f(y_k)}{f(z_k)}$. Under some conditions on H and for nonzero arbitrary values of β, (7) is an optimal method of order four.

Following this idea, we design a class of three-steps methods, that satisfies the following result.

Theorem 1 *Let $\xi \in I$ be a simple zero of a sufficiently differentiable function $f : I \subset \mathbb{R} \to \mathbb{R}$ on the open interval I. Let H and G be sufficiently differentiable real functions and x_0 be an initial approximation close enough to ξ. If $\beta = 1$ and H and G satisfy $H(0) = 1$, $H'(0) = 2b_1$, $H''(0) = 2b_1(2b_1 + b_2)$, $G(0,0) = G_v(0,0) = 1$, $G_u(0,0) = 2b_1$, $G_{uv}(0,0) = 4b_1$ and $G_{uu} = 2b_1(3b_1 + b_2)$, then the method*

$$y_k = x_k - \beta\frac{f(x_k)}{f'(x_k)},$$

$$\begin{aligned} z_k &= y_k - H(u_k)\frac{f(y_k)}{f'(x_k)}, \\ x_{k+1} &= z_k - G(u_k,v_k)\frac{f(z_k)}{f'(x_k)}, \end{aligned} \tag{8}$$

where $u_k = \dfrac{f(y_k)}{b_1 f(x_k) + b_2 f(y_k)}$, *being* b_1 *and* b_2 *arbitrary real parameters and* $v_k = \dfrac{f(z_k)}{f(y_k)}$, *has order of convergence eight for any value* b_1 *different from zero.*

From Theorem 1, we see that the most simple functions H and G are

$$\begin{aligned} H(u) &= 1 + 2b_1u + b_1(2b_1+b_2)u^2, \\ G(u,v) &= 1 + 2b_1u + v + b_1(3b_1+b_2)u^2 + 4b_1uv. \end{aligned} \tag{9}$$

If $b_1 = 1$ and $b_2 = -2$ then $H = 1 + 2u$ and the first two steps of (8) correspond to the Ostrowski's method. In this case function G is $G = 1 + 2u + v + u^2 + 4uv$.

Respect to the class (b) of iterative schemes and from (7), we obtain the following result.

Theorem 2 *Let* $\xi \in I$ *be a simple zero of a sufficiently differentiable function* $f : I \subset \mathbb{R} \to \mathbb{R}$ *on the open interval* I *and be* x_0 *one initial approximation close enough to* ξ. *Let* H *and* G *be sufficiently differentiable real functions which satisfy:* $H(0,0) = H_u(0,0) = H_v(0,0) = 1$, $H_{uu}(0,0) = H_{vv}(0,0) = 2$, $H_{uv}(0,0) = 0$, $G(0,0,0) = G_u(0,0,0) = G_v(0,0,0) = G_w(0,0,0) = 1$, $G_{uu}(0,0,0) = G_{vv}(0,0,0) = G_{uw}(0,0,0) = G_{vw}(0,0,0) = 2$, $G_{uv}(0,0,0) = 1$, $|G_{ww}(0,0,0)| < \infty$. *Then the method*

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[z_k,x_k]}, \\ t_k &= y_k - H(u_k,v_k)\frac{f(y_k)}{f[z_k,x_k]}, \\ x_{k+1} &= t_k - G(u_k,v_k,w_k)\frac{f(t_k)}{f[z_k,x_k]}, \end{aligned} \tag{10}$$

where $z_k = x_k + \beta f(x_k)$, $u_k = \dfrac{f(y_k)}{f(x_k)}$, $v_k = \dfrac{f(y_k)}{f(z_k)}$ *and* $w_k = \dfrac{f(t_k)}{f(y_k)}$, *has order of convergence eight for any value* β *different from zero.*

The most simple functions H and G in this case are

$$\begin{aligned} H(u,v) &= 1 + u + v + u^2 + v^2, \\ G(u,v,w) &= 1 + u + v + w + u^2 + v^2 + uv + 2(vw + uw). \end{aligned} \tag{11}$$

Remark 1 *The families of three-point methods (8) and (10) require four functional evaluations and have order of convergence eight. Therefore, these families are optimal in the sense of the Kung-Traub conjecture and have computational efficiency index* $8^{1/4} \approx 1.6818$.

In the following sections we are going to use the elements of the families (8) and (10) obtained by choosing the weight functions (9) and (11) and the values of the parameters $b_1 = 1$ and $b_2 = 0$. These elements will be denoted by M1 and M2, respectively.

3 Dynamical aspects

The dynamical properties of the rational function associated to an iterative method acting on a polynomial give us important information about numerical features of the method as its stability and reliability (see [15]).

The dynamical behavior of the orbit of a point on the complex plane can be classified depending on its asymptotic behavior. In this way, a point in the Riemann sphere $z_0 \in \hat{\mathbb{C}}$ is a fixed point of R if $R(z_0) = z_0$. A fixed point is attracting, repelling or neutral if $|R'(z_0)|$ is less than, greater than or equal to 1, respectively. Moreover, if $|R'(z_0)| = 0$, the fixed point is superattracting.

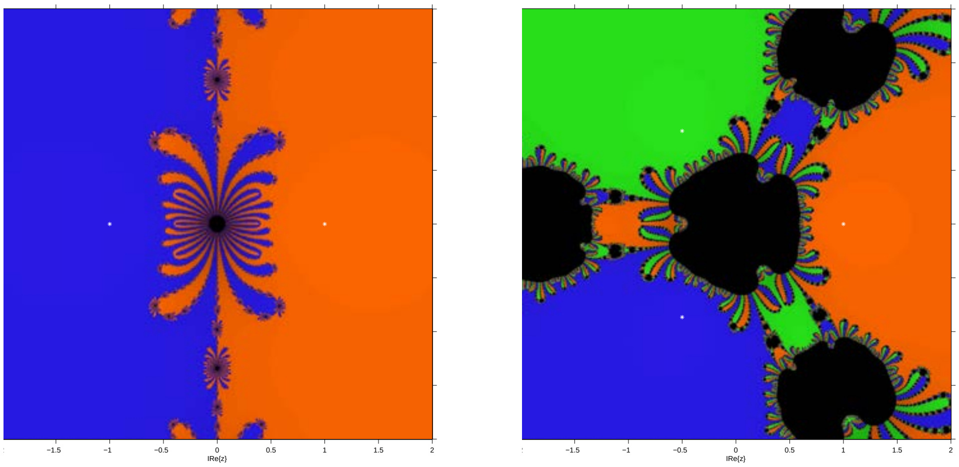
If z^* is an attracting fixed point of the rational function R, its basin of attraction $\mathcal{A}(z^*)$ is defined as the set of pre-images of any order such that

$$\mathcal{A}(z_f^*) = \left\{ z_0 \in \hat{\mathbb{C}} : R^n(z_0) \to z^*, n \to \infty \right\}. \tag{12}$$

The set of points whose orbits tends to an attracting fixed point z_f^* is defined as the Fatou set, $\mathcal{F}(R)$. The complementary set in the Riemann sphere, the Julia set $\mathcal{J}(R)$, is the closure of the set consisting of its repelling fixed points, and establishes the borders between the basins of attraction.

In order to draw the dynamical planes, each point of the complex plane is considered as a starting point of the iterative scheme and it is painted in different colors depending on the point which it has converged to. The figures has been generated for values of z_0 in $[-2, 2] \times [-2, 2]$, with a mesh of 800×800 points and 80 iterations per point. Depending on the number of iterations needed to converge, the color of the starting point will be brighter (less iterations) or darker (more iterations). We will represent the dynamical behavior of the mentioned elements of the two suggested classes, (8) and (10), for low-degree polynomials, showing their stability and the amplitude of the convergence regions in these cases.

When the element of family (8) M1 is considered, we observe in Figure 1 the basin of attraction of the different roots of low-degree polynomials. It is possible to observe some regions of slow convergence, in form of "flowers", whose petals make narrower while the convergence is slower. The black point in the center of the flowers correspond to regions



(a) $x^2 - 1$

(b) $x^3 - 1$

Figure 1: Dynamical planes for the eighth-order family with derivatives

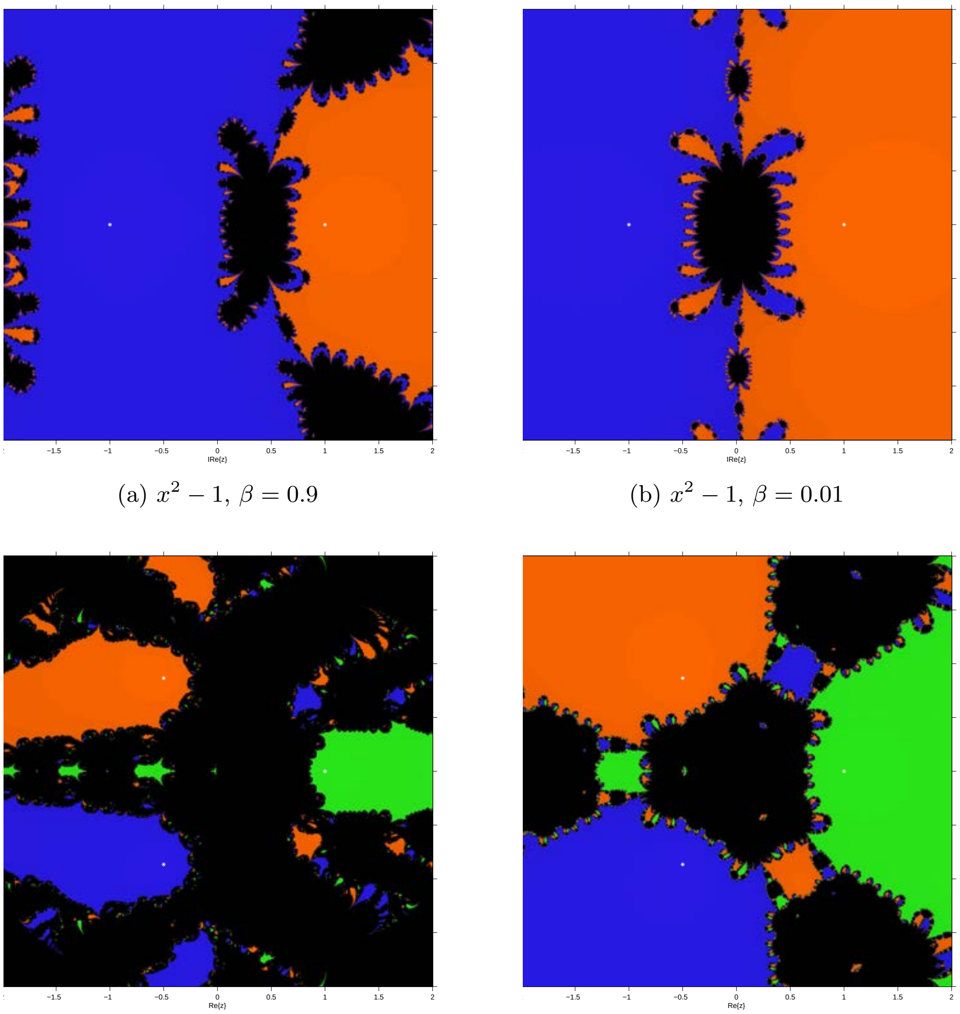
of no convergence. The observed dynamical planes define wide amplitudes for the different basins of the roots.

If we consider now the rational function associated to the element of family (10) and analyze its dynamical behavior, we find that there exist black regions in the dynamical planes (see Figure 2) whose orbits do not tend to any of the roots, but to the infinity. So, the are wide regions of no convergence. Moreover, it can be observed that the amplitude of the convergence regions is narrower in some cases. Nevertheless, the convergence improves when *beta* is closer to zero than when it is near one.

In Figure 3, we show the dynamical planes corresponding to function $f(x) = 10xe^{-x^2} - 1$ in $[-2,2] \times [-2,2]$, where two simple roots appear. These planes have been obtained by using Newton, M1 and M2 methods. In all cases, it can be observed that wide regions of non convergence appear. Some of them are due to poles of the function f. We observe that the behavior of the proposed methods is not very different from the one of Newton's scheme.

4 Numerical results

In this section we demonstrate the convergence behavior of the proposed families (8) and (10). In the numerical test made, variable precision arithmetics has been used, with 4000 digits of mantissa in Matlab R2010a. In our numerical experiments to compare the derivative-free iterative methods and the iterative methods that use derivative, we



(a) x^2-1, $\beta=0.9$

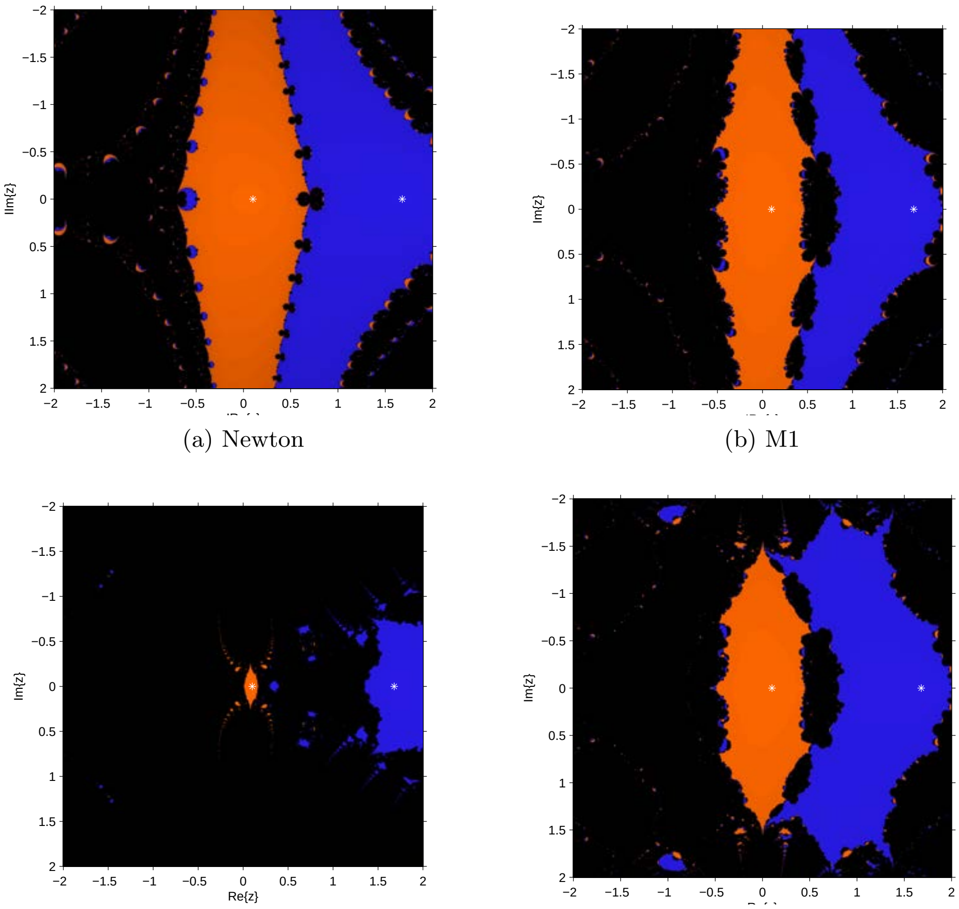
(b) x^2-1, $\beta=0.01$

(c) x^3-1, $\beta=0.9$

(d) x^3-1, $\beta=0.01$

Figure 2: Dynamical planes for the eighth-order family without derivatives

use Newton's method (NM): $x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}$, Steffensen's method (SM): $x_{k+1}=x_k-\frac{f(x_k)^2}{f(x_k+f(x_k))-f(x_k)}$, Ostrowski's method (OM) [17]: $x_{k+1}=y_k-\frac{f(y_k)}{f(x_k)-2f(y_k)}\frac{f(x_k)}{f'(x_k)}$, where $y_k=x_k-\frac{f(x_k)}{f'(x_k)}$, and several iterative methods, quoted below



(a) Newton

(b) M1

(c) M2, $\beta = 0.9$

(d) M2, $\beta = 0.01$

Figure 3: Dynamical planes for function $10xe^{-x^2} - 1$

KWM [4]

$$
\begin{aligned}
y_k &= x_k - \frac{f(x_k)}{f'(x_k)} \\
z_k &= y_k - H_1 \frac{f(y_k)}{f'(x_k)} \\
x_{k+1} &= z_k - \left[(1 + H_1)^2 + (1 + 4H_1) H_2 \right] \frac{f(z_k)}{f'(x_k)}
\end{aligned}
$$

where $H_1 = \frac{f(x_k)}{f(x_k) - 2f(y_k)}$ and $H_2 = \frac{f(z_k)}{f(y_k) - 3f(z_k)}$.

CTVM [2]

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f'(x_k)} \\ z_k &= y_k - H\frac{f(y_k)}{f'(x_k)} \\ x_{k+1} &= u_k - \left(1 + H + \frac{1}{2}T + L\right)\frac{f(z_k)}{f'(x_k)} \end{aligned}$$

where $H = \frac{f(x_k)}{f(x_k)-2f(y_k)}$, $T = \frac{f(z_k)}{f(y_k)-2f(z_k)}$ and $L = 3\frac{u_k-z_k}{y_k-x_k}$.

ZLHM [12]

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[x_k, z_k]} \\ z_k &= y_k - \frac{f(y_k)}{f[y_k, x_k] + f[y_k, x_k, z_k](y_k - x_k)} \\ x_{k+1} &= u_k - \frac{f(u_k)}{f[u_k, y_k] + f[u_k, y_k, x_k](u_k - y_k) + f[u_k, y_k, x_k, z_k](u_k - y_k)(u_k - x_k)} \end{aligned}$$

SKM [10]

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f[x_k, z_k]} \\ u_k &= y_k - H\frac{f(x_k)}{f[x_k, z_k]} \\ x_{k+1} &= u_k - G\frac{f(u_k)}{f[x_k, z_k]} \end{aligned}$$

where $H = 1 + w_1(x_k) + w_2(x_k)$, $w_1(x_k) = \frac{f(y_k)}{f(x_k)}$, $w_2(x_k) = \frac{f(y_k)}{f(z_k)}$, $G = g_1 + g_2 + g_3$, $g_1 = 1 + (2 - f[x_k, z_k])w_1(x_k) + (1 - f[x_k, z_k])w_1(x_k)^2$, $g_2 = (4 - f[x_k, z_k](6 + f[x_k, z_k](-4 + f[x_k, z_k])))w_1(x_k)^3$, $g_3 = w_3(x_k) + w_3(x_k)^2 + (4 - 2f[x_k, z_k])w_4(x_k)$, $w_3(x_k) = \frac{f(u_k)}{f(y_k)}$ and $w_4 = \frac{f(u_k)}{f(z_k)}$.

Table 1 shows the expression of the test functions, the roots with sixteen significant digits and the initial approximation x_0 which is the same for all methods. Displayed in Table 2 is the number of iterations (It), the computational order of convergence (COC), the absolute values of the function $|f(x_k)|$ and root obtained at each method ξ.

Table 1: Test functions and their roots

Test functions	Roots	Starting points
$f_1 = \sqrt{x^4+8}\ \sin\left(\frac{\pi}{x^2+2}\right) + \frac{x^3}{x^4+1} - \sqrt{6} + \frac{8}{17}$	$\xi_1 = -2.000000000000000$ $\xi_2 = -1.149212674609088$	$x_0 = -1.8$
$f_2 = x\exp(x^2) - \sin(x^2) + 3\cos(x) + 5$	$\xi = -1.201576112092293$	$x_0 = 2$
$f_3 = \sqrt{x^2+2x+5} - 2\sin(x) - x^2 + 3$	$\xi_1 = +2.3319676558839640$ $\xi_2 = -2.573166514902827$	$x_0 = 1$
$f_4 = x^4 + \sin\left(\frac{\pi}{x^2}\right) - 5$	$\xi_1 = +1.414213562373095$ $\xi_2 = -1.414213562373095$	$x_0 = 2$
$f_5 = \left(\sin(x) - \frac{x}{2}\right)^2$	$\xi = 0$ (double root)	$x_0 = 0.5$

5 Conclusions

These results that shown in the Table 1 are in accordance with the efficiency analysis the show in Theorem 1 and Theorem 2 presented in Section 2. Both the numerical results, such as the dynamic analysis show that derivative-free methods are more sensitive to the initial approximation comparing them with those that use derivatives.

Acknowledgements

This research was supported by Ministerio de Ciencia y Tecnología MTM2011-28636-C02-02 and FONDOCYT 2011-1-B1-33 República Dominicana.

References

[1] S. Artidiello, F. Chicharro, A. Cordero and J. R. Torregrosa, *Local convergence and dynamical analysis of a new family of optimal fourth-order iterative methods*, Inter. J. of Comp. Math., DOI: 10.1080/00207160.2012.748900.

[2] A. Cordero, J.R. Torregrosa and M. Vassileva, *Tree-step iterative method with a optimal eighth order of convergence*, J. of Comp. and Appl. Math, **235** (2011) 3189–3194.

Table 2: Comparison of various derivative-free iterative methods

		NM	SM	OM	KWM	CTVM	M1	ZLHM	SKM	M2
f_1	It	9	9	5	3	3	4	3	4	3
	COC	2.0000	2.0000	4.0000	8.2176	8.3082	8.0000	8.1662	7.9999	8.0000
	$\|f(x_n)\|$	1.6e-479	2.3e-574	8.8e-752	9.8e-359	1.4e-329	9.6e-2525	6.6e-431	4.2e-1882	9.6e-2522
	root	ξ_1	ξ	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1
f_2	It	14	n.c.	5	3	3	4	4	n.c.	4
	COC	2.0000		4.0000	7.9547	7.9386	8.0203	8.0038		8.0203
	$\|f(x_n)\|$	2.6e-553		7.9e-1063	3.5e-517	8.6e-485	5.0e-528	3.7e-1066		5.0e-528
	root	ξ		ξ	ξ	ξ	ξ	ξ		ξ
f_3	It	8	8	5	3	3	3	3	14	3
	COC	2.0000	2.0000	3.9998	8.0286	7.9075	7.9753	8.0003	7.6578	7.9753
	$\|f(x_n)\|$	3.2e-422	1.8e-334	2.7e-1292	5.6e-560	6.7e-622	4.0e-545	9.0e-556	4.4e-080	4.0e-545
	root	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1	ξ_1
f_4	It	9	70	5	4	3	n.c.	4	6	n.c.
	COC	2.0000	2.0000	4.0000	8.0232	7.6465		8.0353	8.0000	
	$\|f(x_n)\|$	8.2e-425	5.1e-306	5.6e-621	1.3e-1714	2.3e-333		1.3e-092	2.7e-1750	
	root	ξ_1	ξ_2	ξ_2	ξ_1	ξ_1		ξ_1	ξ_1	
f_5	It	536	536	268	n.c.	194	226	179	226	225
	COC	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000
	$\|f(x_n)\|$	7.1e-325	7.9e-325	1.0e-326		3.1e-325	3.5e-325	2.5e-325	2.0e-324	3.5e-325
	root	ξ	ξ	ξ	ξ	ξ	ξ	ξ	ξ	ξ

[3] Y. Khan, M. Fardi and K. Sayevand, *A new general eighth-order family of iterative methods for solving nonlinear equations*, Appl. Math. L. **25** (2012) 2262–2266.

[4] J. Kou and X. Wang, *Some improvements of Ostrowski's method*, Appl. Math. L., **23** (2010) 92–96.

[5] Z. Liu, Q. Zheng and P. Zhao, *A variant of Steffensen's method of fourth-order convergence and its applications*, Appl. Math. and Comput., **216** (2010) 1978–1983.

[6] J. M. Ortega and W. G. Rheinboldt, *Iterative solutions of nonlinear equations in several variables*, Academic Press, New York, 1970.

[7] M.S. Petković, S. Ilić and J. Džunić, *Derivative free two-point methods with and without memory for solving nonlinear equations*, Appl. Math. and Comput., **217** (2010) 1887–1895.

[8] M. Petkovic, B. Neta, L. Petkovic and J. Dzunic, *Multipoint methods for solving nonlinear equations*, Academic Press, 2012.

[9] H. Ren, Q. Wu and W. Bi, *A class of two-step Steffensen type methods with fourth-order convergence*, Appl. Math. and Comput., **209** (2009) 206–210.

[10] F. Soleymani and S.K. Khattri, *Finding simple roots by seventh- and eighth-order derivative-free methods*, Int. J. Math. Models Meth. Appl. Sci., **1** (2012) 45–52.

[11] J. F. Traub, *Iterative methods for the solution of equations*, Chelsea Publishing Company, New York, 1977.

[12] Q. Zheng, J. Li and F. Huang, *An optimal Steffensen-type family for solving nonlinear equations*,Appl. Math. and Comput. **217**, (2011) 9592–9597.

[13] H. T. Kung and J. F. Traub, *Optimal order of one-point and multipoint iteration*, J. Assos. Comput. Math. **21** (1974) 643–651.

[14] Q. Zheng, P. Zhao, L. Zhang and W. Ma, *Variants of Steffensen-secant method and applications*, Appl. Math. and Comput., **216** (2010) 3486–3496.

[15] P. Blanchard, *The Dynamics of Newton's Method*, Proc. of Symposia in Applied Math. **49** (1994) 139–154.

[16] S. Weerakoon and T.G.I. Fernando, *A variant of Newton's method with accelerated third-order convergence*, Appl. Math. L., **13** (2000) 87–93.

[17] A.M. Ostrowski, Solution of Equations and System of Equations, Prentice-Hall, Englewood Cliffs, NJ, USA, 1964.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

On $\mathbb{Z}_2\mathbb{Z}_2[u]-$Additive Codes

Ismail Aydogdu[1], Taher Abualrub[2] and Irfan Siap[1]

[1] *Department of Mathematics, Yildiz Technical University*

[2] *Department of Mathematics and Statistics, American University of Sharjah*

emails: `iaydogdu@yildiz.edu.tr`, `abualrub@aus.edu`, `isiap@yildiz.edu.tr`

Abstract

In this paper, we introduce a new class of additive codes called $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive codes. This is a generalization towards another direction of recently introduced $\mathbb{Z}_2\mathbb{Z}_4-$additive codes. We determine the generator and parity-check matrices. We introduce a Gray map and give some examples of optimal codes which are the binary Gray images of $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive codes.

Key words: $\mathbb{Z}_2\mathbb{Z}_2[u]-$Additive Codes, Generator matrix, Parity-check matrix
MSC 2000: AMS codes (optional)

1 Introduction

Codes over rings with the remarkable paper by Hammons at el.[4] has been studied intensively for the last four decades. $\mathbb{Z}_2\mathbb{Z}_4-$additive codes, which are the generalization of binary linear codes and quaternary linear codes, have attracted many researchers in algebraic coding theory recently. The structure of $\mathbb{Z}_2\mathbb{Z}_4-$additive codes and their duals has been determined [2, 3]. Aydogdu and Siap generalize these codes as $\mathbb{Z}_2\mathbb{Z}_2^s$-additive codes [1]. Another important ring of four elements is the the ring $R = \mathbb{Z}_2 + u\mathbb{Z}_2 = \{0, 1, u, 1+u\}$ where $u^2 = 0$. A linear code $\mathscr{C}$ over R is permutation equivalent to a code with generator matrix,

$$G = \begin{bmatrix} I_{k_1} & A & B_1 + uB_2 \\ 0 & uI_{k_2} & uD \end{bmatrix}$$

where A, B_1, B_2 and D are matrices over $\mathbb{Z}_2$.

We know that the ring $\mathbb{Z}_2$ is a subring of the ring R. Being inspired by the structure of $\mathbb{Z}_2\mathbb{Z}_4-$additive codes, we introduce the ring,

$$\mathbb{Z}_2\mathbb{Z}_2[u] = \{(a,b) \mid a \in \mathbb{Z}_2 \text{ and } b \in R\}.$$

The ring $\mathbb{Z}_2\mathbb{Z}_2[u]$ is not well-defined with respect to the usual multiplication by $u\in R$. Therefore, the ring is not an $R-$module. To make it well-defined and enrich with an algebraic structure we introduce a new multiplication as follows.

We define a map

$$\eta : R\to\mathbb{Z}_2$$
$$\eta(a+bq)=a$$

as $\eta(0)=0,\eta(1)=1,\eta(u)=0$ and $\eta(u+1)=1$. It is clear that η is a ring homomorphism. Using this map, we define a scalar multiplication as follows. For $v=(a_0,a_1,...,a_{\alpha-1,}b_0,b_1,...,b_{\beta-1})\in\mathbb{Z}_2^\alpha\times(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta$ and $d\in R$ we have

$$dv=\left(\eta(d)a_0,\eta(d)a_1,...,\eta(d)a_{\alpha-1,}db_0,db_1,...,db_{\beta-1}\right). \quad (*)$$

2 $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive Codes

In this section, we introduce $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive codes. We determine their generators. We also define a dual code and obtain its parity check matrix.

Definition A linear code $\mathscr{C}$ is called a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code if it is a $\mathbb{Z}_2+u\mathbb{Z}_2-$submodule of $\mathbb{Z}_2^\alpha\times(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta$ with respect to the scalar multiplication defined in (*). Then the binary image $\Phi(\mathscr{C})=C$ is called $\mathbb{Z}_2\mathbb{Z}_2[u]-$linear code of length $n=\alpha+2\beta$ where Φ is map from $\mathbb{Z}_2^\alpha\times(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta$ to $\mathbb{Z}_2^n$ defined as;

$$\Phi(a,b)=\left(a_0,a_1,...,a_{\alpha-1,}\phi(b_0),\phi(b_1),...,\phi(b_{\beta-1})\right)$$

for all $a=(a_0,a_1,...,a_{\alpha-1})\in\mathbb{Z}_2^\alpha$ and $b=(b_0,b_1,...,b_{\beta-1})\in(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta$ and $\phi:R\to\mathbb{Z}_2^2$ is defined by $\phi(0)=(0,0)$, $\phi(1)=(0,1),\phi(u)=(1,1)$, $\phi(u+1)=(1,0)$. According to the definition, we can say that a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code is isomorphic to an abelian structure like $\mathbb{Z}_2^{k_0}\times\mathbb{Z}_2^{2k_1}\times\mathbb{Z}_2^{k_2}$. Here, the first α coordinates of $\mathscr{C}$ are elements from $\mathbb{Z}_2$ and remaining β coordinates are from $\mathbb{Z}_2+u\mathbb{Z}_2$. Let $\mathscr{C}_\beta^F$ be the submodule, $\mathscr{C}_\beta^F=\{(a,b)\in\mathbb{Z}_2^\alpha\times(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta \mid b \text{ free over } (\mathbb{Z}_2+u\mathbb{Z}_2)^\beta\}$ and we say $dim(\mathscr{C}_\beta^F)=k_1$. Let $D=\mathscr{C}\backslash\mathscr{C}_\beta^F=\mathscr{C}_0\oplus\mathscr{C}_1$ such that $\mathscr{C}_0=\{(a,b)\in\mathbb{Z}_2^\alpha\times(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta \mid a\neq0\}\subseteq\mathscr{C}\backslash\mathscr{C}_\beta^F$ and $\mathscr{C}_1=\{(a,b)\in\mathbb{Z}_2^\alpha\times(\mathbb{Z}_2+u\mathbb{Z}_2)^\beta \mid a=0\}\subseteq\mathscr{C}\backslash\mathscr{C}_\beta^F$. Now, denote the dimension of $\mathscr{C}_0$ as a k_0 and denote the dimension of $\mathscr{C}_1$ as a k_2. Considering all these parameters we say $\mathscr{C}$ is of type $(\alpha,\beta;k_0,k_1,k_2)$.

2.1 Generator Matrices of $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive Codes

Theorem 2.1.1 Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(\alpha,\beta;k_0,k_1,k_2)$. Then $\mathscr{C}$ is permutation equivalent to a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code with the standard form matrix,

$$G=\begin{bmatrix} I_{k_0} & A_1 & 0 & 0 & uT \\ 0 & S & I_{k_1} & A & B_1+uB_2 \\ 0 & 0 & 0 & uI_{k_2} & uD \end{bmatrix} \quad (1)$$

where A, A_1, B_1, B_2, D, S and T are matrices over $\mathbb{Z}_2$.

Example 2.1.2 Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(2,3;1,2,0)$ with generator matrix,

$$\begin{bmatrix} 1 & 1 & u & u & u \\ 0 & 1 & 1 & 0 & 1+u \\ 1 & 0 & 1 & u & u \end{bmatrix}. \tag{2}$$

Then $\mathscr{C}$ has the standard form matrix as follows,

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 & u \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and $\mathscr{C}$ has $|\mathscr{C}| = 2^1 4^2 = 32$ codewords.

2.2 Duality of $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive Codes

We define an inner product for $v, w \in \mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta$ as

$$\langle v, w \rangle = u \left(\sum_{i=1}^{\alpha} v_i w_i \right) + \sum_{j=\alpha+1}^{\alpha+\beta} v_j w_j \in (\mathbb{Z}_2 + u\mathbb{Z}_2).$$

Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code. The additive dual of $\mathscr{C}$, denoted by $\mathscr{C}^\perp$, is then defined in standard way as

$$\mathscr{C}^\perp = \left\{ w \in \mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta \mid \langle v, w \rangle = 0 \; for \; all \; v \in \mathscr{C} \right\}.$$

We further define the following auxiliary mappings for constructing an additive dual code of a $\mathbb{Z}_2(\mathbb{Z}_2 + u\mathbb{Z}_2)-$additive code. Let $\chi : \mathbb{Z}_2 \to (\mathbb{Z}_2 + u\mathbb{Z}_2)$ such that $\chi(0) = 0$ and $\chi(1) = u$. And let $\psi : (\mathbb{Z}_2 + u\mathbb{Z}_2) \to \mathbb{Z}_2$ for all $x \in (\mathbb{Z}_2 + u\mathbb{Z}_2)$ as;

$$\psi(x) = \begin{cases} 0, x = 0 \text{ or } x = u \\ 1, x = 1 \text{ or } x = u+1 \end{cases}.$$

Finally, we define also the identity map, $\iota : \mathbb{Z}_2 \to (\mathbb{Z}_2 + u\mathbb{Z}_2)$ such that $\iota(0) = 0$ and $\iota(1) = 1$. Also, denote the extensions of these functions with the same notations. Therefore, we have; (χ, I_d) : $\mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta \to (\mathbb{Z}_2 + u\mathbb{Z}_2)^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta$, $(\psi, I_d) : (\mathbb{Z}_2 + u\mathbb{Z}_2)^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta \to \mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta$ and $(\iota, I_d) : \mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta \to (\mathbb{Z}_2 + u\mathbb{Z}_2)^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta$. The following lemma presents an intersection between these maps.

Lemma 2.2.1 Let $v \in \mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta$ and $w \in (\mathbb{Z}_2 + u\mathbb{Z}_2)^{\alpha+\beta}$ then $\langle \chi(v), w \rangle_u = \langle v, \psi(w) \rangle$, where $\langle \, , \, \rangle_u$ denotes the standard inner product of vectors in $(\mathbb{Z}_2 + u\mathbb{Z}_2)$.

Corollary 2.2.2 If $v, w \in \mathbb{Z}_2^\alpha \times (\mathbb{Z}_2 + u\mathbb{Z}_2)^\beta$ then $\langle \chi(v), \iota(w) \rangle_u = \langle v, w \rangle$.

Proposition 2.2.3 Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(\alpha, \beta; k_0, k_1, k_2)$. Then $\mathscr{C}^\perp = \psi(\chi(\mathscr{C})^\perp)$.

2.3 Parity-check Matrices of $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive Codes

As in classical case, the generator matrix of the dual code is important. In the following theorem, we give the standard form of the generator matrix of the dual code.

Theorem 2.3.1 Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(\alpha,\beta;k_0,k_1,k_2)$ with standard form matrix (1). Then the generator matrix for the additive dual code $\mathscr{C}^{\perp}$ is given by

$$H=\begin{bmatrix} -A_1^t & I_{\alpha-k_0} & -uS^t & 0 & 0 \\ -T^t & 0 & -(B_1+uB_2)^t+D^tA^t & -D^t & I_{\beta-k_1-k_2} \\ 0 & 0 & -uA^t & uI_{k_2} & 0 \end{bmatrix}. \tag{3}$$

Example 2.3.2 Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(2,3;1,1,1)$ with the generator matrix (2). Then we can write the parity-check matrix of $\mathscr{C}$ as follows,

$$H=\begin{bmatrix} 1 & 1 & u & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

It is clear that $\mathscr{C}$ is of type $(2,3;1,1,0)$ and has $|\mathscr{C}|=2^14^1=8$ codewords.

3 Examples of Optimal Codes

We present some optimal codes derived from this family of additive codes.

Example 3.1 Let $\mathscr{C}$ be a $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(7,7;3,1,0)$ with the generator matrix

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1+u & 1 & 1+u & 1+u & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & u & u & u & 0 & u & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & u & u & u & 0 & u & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & u & u & u & 0 & u \end{bmatrix}.$$

Using the gray map which we defined before, we have $\Phi(\mathscr{C})=C$ is a $[21,5,10]$ binary code.

Example 3.2 Consider the $\mathbb{Z}_2\mathbb{Z}_2[u]-$additive code of type $(9,9;6,1,0)$ with the generator matrix

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1+u & 1+u & 1+u & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & u & 0 & 0 & u & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & u & 0 & 0 & u & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & u & 0 & 0 & u & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & u & 0 & 0 & u & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & u & 0 & 0 & u & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & u & 0 & 0 & u \end{bmatrix}.$$

Then the binary image of this code has parameters $[27,8,10]$.

Conclusion

In this work, we introduce linear codes that are special modules and a Gray map that maps this family to linear binary codes. We study their generator matrices and by introducing an inner product we define their duals and determine their parity-check matrices. We conclude by presenting some optimal binary codes obtained through this family. Since this is a new definition and a promising family, this topic awaits a deeper exploration.

References

[1] I. Aydogdu and I. Siap, "The Structure of $\mathbb{Z}_2\mathbb{Z}_2^s-$ Additive Codes: Bounds on the minimum distance", Applied Mathematics and Information Sciences(AMIS), accepted (2013).

[2] M. Bilal, J. Borges, S.T. Dougherty, C. Fernández-Córdoba, "Optimal Codes Over $\mathbb{Z}_2 \times \mathbb{Z}_4$", VII Jornadas de Matemática Discreta y Algor ítmica Castro Urdiales, 7-9 de julio de 2010.

[3] J. Borges, C. Fernández-Córdoba, J. Pujol, J. Rif́a and M. Villanueva, "$\mathbb{Z}_2\mathbb{Z}_4$-linearcodes: Generator Matrices and Duality", Designs, Codes and Cryptography, vol. 54(2), pp. 167-179, 2010.

[4] A.R.Hammons, P.V.Kumar, A.R.Calderbank, N.J.A.Sloane, P.Sole, "The Z4-linearity of kerdock, preparata, goethals and related codes", IEEE Trans. Inform. Theory 40, 301319 (1994).

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Drug Delivery from an ocular implant into the vitreous chamber of the eye

E. Azhdari[1], J.A. Ferreira[1], P. de Oliveira[1] and P.M. da Silva[2]

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

[2] *Department of Physics and Mathematics, Coimbra Institute of Engineering - Portugal*

emails: ebrahim@mat.uc.pt, ferreira@mat.uc.pt, poliveir@mat.uc.pt, pascals@isec.pt

Abstract

A mathematical model which simulates drug delivery from an ocular implant into the vitreous chamber of the eye is proposed. The model consists of coupled systems of partial differential equations linked by interface conditions. The chemical structure, the viscoelastic properties and the diffusion phenomena are taken into account to simulate the evolution of released drug. Numerical simulations that illustrate the interplay between these phenomena are included.

Key words: diffusion-reaction equation, drug delivery, biodegradable implant.

1 Introduction

The vitreous humor is the clear gel that fills the space between the lens and the retina of the eyeball of humans and other vertebrates (Figure 1). It is bounded by the lens, the hyaloid membrane, and the retina. The vitreous is stuck to the retina, but with aging, the vitreous can separate from it. The vitreous humor makes up 80% of the eye to hold its fairly spherical shape.

There are a number of severe diseases that can affect the vitreous and the retina, which must be treated over long periods of time and where drugs must be maintained in their therapeutic windows. Among these diseases we mention:

- Age-related macular degeneration which is a medical condition that usually affects older adults and results in a loss of vision in the center of the visual field because of damage to the retina.

- Glaucoma that is an eye disease in which the optic nerve is damaged and that is normally associated with increased fluid pressure in the anterior chamber of the eye.

- Diabetic retinopathy which is a retinopathy caused by complications of diabetes, that affect the blood vessels of the retina.

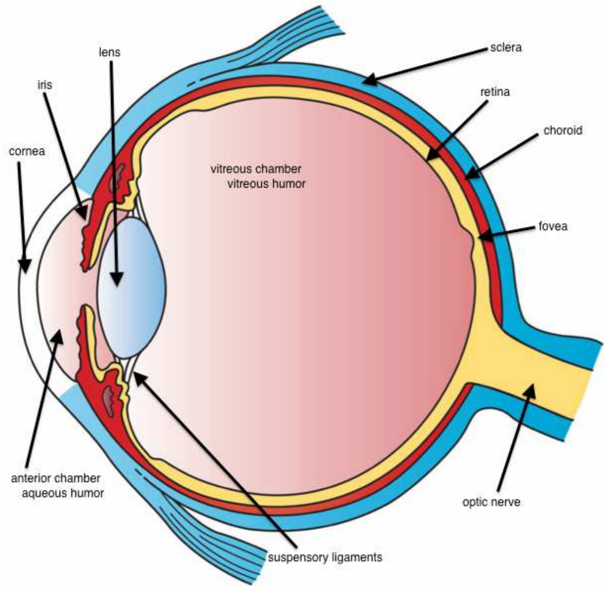


Figure 1: Anatomy of the human eye (http://en.wikipedia.org/wiki/).

Delivering drugs to the vitreous chamber of the eye assumes a crucial role and is a challenging problem due to the presence of various physiological and anatomical barriers. Classical ocular drug delivery systems for posterior segment diseases fall under one of the following categories:

- Systemic delivery: systemic administration of drugs to the blood stream directly, in the form of injections, or by absorption into the blood stream, in the form of pills.

- Topical delivery: topical delivery in the form of ophthalmic drops which is the most common method used to treat ocular diseases.

None of the these first drug delivery systems is effective. In fact systemic delivery is not effective because as the eye has a relatively small size the drug concentration carried by the blood stream is not enough which means that it does not reach the therapeutic window of the drug; with topical delivery just a small fraction of drug reaches the posterior segment of the eye due to physiological barriers which difficult the inlet in vitreous of drug in systemic circulation.

Those classical drug delivery systems are being replaced by direct intravitreal injection or intravitreal implants of drug. As vitreal injections imply several treatments and can cause side effects as the increase of intraocular pressure and the damage of crystalline lens intravitreal implants have deserved much attention these last years. In this paper we will study intravitreal delivery of drug through implants. This delivery system is nowadays the most used method to treat the vitreous chamber of the eye. Intravitreal implants include nonbiodegradable and biodegradable polymers. Controlled release of a therapeutic agent from a biodegradable polymeric system presents an alternative to traditional treatment strategies that can overcome some of the problems associated with pulsed delivery ([9]). Many drugs have a narrow concentration window of effectiveness and may be toxic at higher concentration ([9]), so the ability to predict local drug concentrations is necessary for proper loading of the delivery system. Mathematical models play a central role in drug release because not only they explain the kinetics of the delivery by describing the interplay of the different phenomena as they quantify the effect of physical parameters in the delivery trend. Several authors have studied mathematical models to describe transport and elimination of drugs in the vitreous ([2, 3, 7, 9]). However at the best of our knowledge the delivery of drug from a biodegradable implant has not yet been addressed. There are several type of implants already approved for ophthalmic applications. The model presented in this paper is a general transport model in a biodegradable viscoelastic material. However in the numerical simulations some physical data from a FDA [1] approved intravitreal implant, Ozurdex, have been used (Figure 2) .

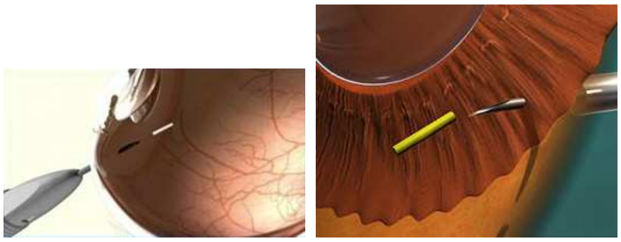


Figure 2: Ocular implant (http://www.tanner-eyes.co.uk/patient_ozurdex.html and http://marcelohosoume.blogspot.pt/2010/10/iluvien-and-future-of-ophthalmic-drug.html).

In Section 2 the geometry of the vitreous chamber of the eye and of the intravitreal implant are described. In Section 3 the mathematical model is presented. Numerical simulations that illustrate the kinetics of the drug release and emphasis the effect of degradation and viscoelasticity are exhibited in Section 4. Finally in Section 5 some conclusions are

[1] FDA- Food and Drug Administration.

addressed.

2 Geometry

The geometrical model of the human eye adopted in the present study is shown in Figure 3 and is based on the physiological dimensions ([7]).

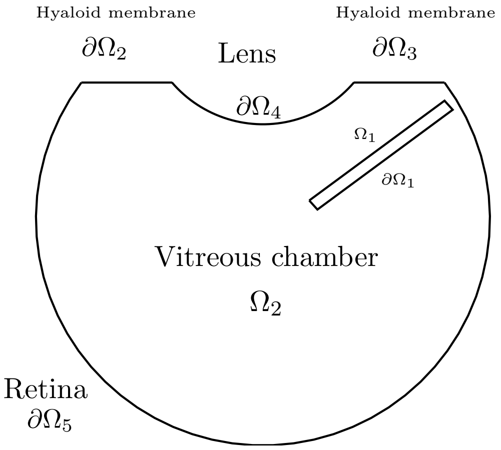


Figure 3: Geometry of the vitreous chamber of the human eye (Ω_2), hyaloid membrane ($\partial\Omega_2$, $\partial\Omega_3$), lens ($\partial\Omega_4$), retina ($\partial\Omega_5$), ocular implant (Ω_1) and its boundary ($\partial\Omega_1$).

The vitreous chamber is mainly composed of vitreous humor and comprises about two-third of the eye. The lens is located behind the iris and is modeled here as an ellipsoid. The hyaloid membrane and the lens separate the anterior chamber and the posterior chamber of the eye from the vitreous chamber. The retina forms the boundary for the vitreous on the posterior surface and is modeled as a spherical surface with a radius of 9.1 mm. The intravitreal implant is placed into the vitreous, as shown in Figure 3, and it is geometrically represented by a cylinder with radius 0.023 mm and height 0.6 mm.

3 Mathematical model

The implant with dispersed drug is placed into the vitreous, near the retina (Figure 3). The drug is released in a controlled manner through the vitreous which is a porous media, and its target is the retina, affected by an inflammatory process.

The diffusion-reaction equation that describes the drug dynamics in the polymer is represented by

$$\begin{cases} \dfrac{\partial C_1}{\partial t} = \nabla(D_1(M)\nabla C_1) + D_v\Delta\sigma - k_1 C_1 \texttt{ in } \Omega_1 \times (0,T] \\[2ex] \dfrac{\partial \sigma}{\partial t} + \dfrac{E}{\mu}\sigma = EC_1 \texttt{ in } \Omega_1 \times (0,T] \\[2ex] \dfrac{\partial M}{\partial t} + \beta_1 M = \beta_2 C_1 \texttt{ in } \Omega_1 \times (0,T] \end{cases}, \tag{1}$$

where Ω_1 stands for the implant (a cylindrical device with dispersed drug), C_1 is the drug concentration in the polymer, σ is the stress exerted by the polymer and M represents the polymer molecular weight. The diffusion coefficient of the drug in the polymer, D_1, is a function of the molecular weight, the parameter D_v stands for a stress driven diffusion coefficient and k_1 is the degradation rate of the drug. The second equation in (1) results from the Maxwell fluid model ([4],)

$$\frac{\partial \sigma}{\partial t} + \frac{E}{\mu}\sigma = E\frac{\partial \varepsilon}{\partial t}, \tag{2}$$

that relates the stress with the strain ε where E represents the Young modulus of the polymer and μ its viscosity. To eliminate the strain from the system we consider that the strain satisfies

$$\varepsilon = k\int_0^t C_1(x,s)ds, \tag{3}$$

where k is a dimensional constant. From (2) and (3) the second equation in (1) is obtained where the constant k has been absorbed by E.

The third equation in (1) represents the degradation of the polymer and β_1, β_2 are physical constant which are material dependent. It is expected that as the polymer erodes, the diffusion of the drug concentration becomes larger, so we define

$$D_1(M) = \lambda e^{\frac{M_0}{M_0+M}}, \tag{4}$$

where M_0 is the initial molecular weight.

Let us now consider the drug dynamics in the vitreous, where the diffusion of drug occurs from the polymer towards the vitreous and the retina. In general, mass transport in the vitreous is not described by diffusion only, but convection is equally important. Convection is due to the steady permeation of the aqueous humor through the vitreous, and diffusion is driven by the concentration gradient ([**?**]). To simulate the dynamics of drug in the vitreous we take into account a diffusion reaction equation, where the velocity of aqueous permeation ([2],[**?**],[1],[6]) is given by Darcy's law in Ω_2, as follows:

$$\frac{\partial C_2}{\partial t} + \mathbf{v}.\nabla C_2 - D_2 \Delta C_2 = 0 \, \texttt{in} \; \Omega_2 \times (0,T], \tag{5}$$

and

$$\left\{ \begin{array}{c} \mathbf{v} = -\dfrac{K}{\mu_1} \nabla p \, \texttt{in} \; \Omega_2 \times (0,T] \\ \nabla.\mathbf{v} = 0 \, \texttt{in} \; \Omega_2 \times (0,T] \end{array} \right. . \tag{6}$$

In equation (5) C_2 represents the concentration of the drug in the vitreous, D_2 is the diffusion coefficient of the drug in the vitreous and $\mathbf{v}$ the velocity of aqueous permeation given by (6). In this last system K is the permeability of the vitreous and μ_1 is the viscosity of the permeating aqueous humour ([?]). The term $\dfrac{K}{\mu_1}$ is referred to as the hydraulic conductivity.

Equations (1-6) are completed with initial conditions represented by

$$\left\{ \begin{array}{l} C_1 = c_0, \; \texttt{in} \; \Omega_1, \; t = 0 \\ \sigma = \sigma_0, \; \texttt{in} \; \Omega_1, \; t = 0 \\ M = M_0, \; \texttt{in} \; \Omega_1, \; t = 0 \\ C_2 = 0, \; \texttt{in} \; \Omega_2, \; t = 0 \\ \mathbf{v} = 0, \; \texttt{in} \; \Omega_2, \; t = 0 \\ p = 2000, \; \texttt{in} \; \Omega_2, \; t = 0 \end{array} \right. . \tag{7}$$

Boundary conditions of different type will be used in the model:

- Boundary conditions for the pressure:

  $$p = 2000, \; \texttt{in} \; \partial\Omega_2, \cup \partial\Omega_3 \, t > 0,$$

  $$p = 1200, \; \texttt{in} \; \partial\Omega_5, \; t > 0.$$

  We note that $\partial\Omega_2, \cup \partial\Omega_3$ represent the hyaloid membrane and $\partial\Omega_5$ represents the retina.

- Interface boundary conditions for the flux of drug concentration:

  $$D \nabla C_1.\eta = A(C_1 - C_2), \; \texttt{in} \; \partial\Omega_1, \; t > 0.$$

- Wall conditions for the velocity: $\mathbf{v} = 0$, in the boundary $\partial\Omega_4$, of the vitreous chamber Ω_2, and in the boundary of the implant $\partial\Omega_1$ (Figure 3).

4 Numerical simulations

In this section we will present some simulations to illustrate the behaviour of drug concentration in the implant and in the vitreous. In the case the values of the constants were not available, we used values that make physical sense but that may not correspond to the characteristics of the intravitreous implant. Consequently this study has mainly a qualitative character.

The numerical simulations have been obtained with $C_0 = 1.7887\times10^{-6}$, $M_0 = 0.5\times10^{-6}$ and $\sigma_0 = 0.5 \times 10^{-6}$ (mol/mm^3), where σ_0 represents the initial stress in the implant.

The diffusion coefficient of the drug in the implant is defined considering $\lambda = 1 \times 10^{-11}$ in (4) and its diffusion coefficient in the vitreous is defined by $D_2 = 1\times10^{-8}$. We recall that the diffusion in the polymer will increase as the molecular weight decreases. The following values for the parameters:

$$k_1 = 1 \times 10^{-10},\ \beta_1 = 5 \times 10^{-4},\ \beta_2 = 1 \times 10^{-9},\ \mu = 2 \times 10^{-4},\ E = 1 \times 10^{-7},$$

and

$$A_c = 5 \times 10^{-5},\ D_v = 1 \times 10^{-11},\ \mu_1 = 0.7,\ \ \rho = 970,\ \ K = 0.7 \times 8.4 \times 10^{-8},$$

have been considered. The units of the previous parameters are such that the equations are dimensionally correct when concentrations are considered in mol/mm^3 as previously indicated.

In Figure 4 the drug concentration at time $t = 5\,min$ and $t = 1\,h$ are presented. It can be observed that as time evolves the drug is released and less drug concentration is inside the implant. We remark that the maximum concentration for $t = 5\,min$ is higher than the maximum concentration at $t = 1\,h$.

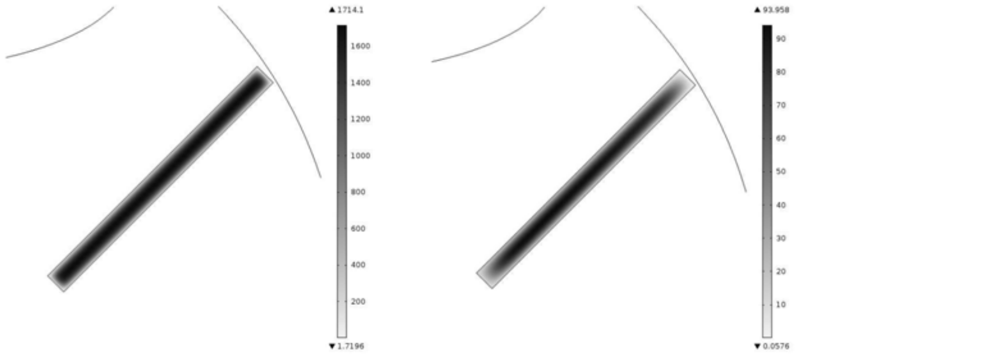


Figure 4: Drug concentration in the implant at $5\,min$ (left) and $1\,h$ (right).

The pressure in the vitreous chamber is showed in Figure 5. The evolution of the pressure from the top ($p = 2000\,Pa$) until the boundary of the vitreous chamber that is in contact with the retina ($p = 1200\,Pa$), can be observed. In Figure 6 the drug concentration in the vitreous chamber is plotted for $t = 5\,min$ and $t = 1\,h$.

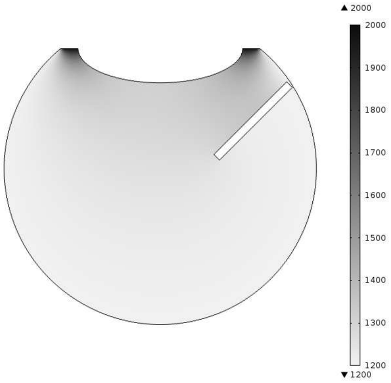


Figure 5: Steady pressure in the vitreous chamber.

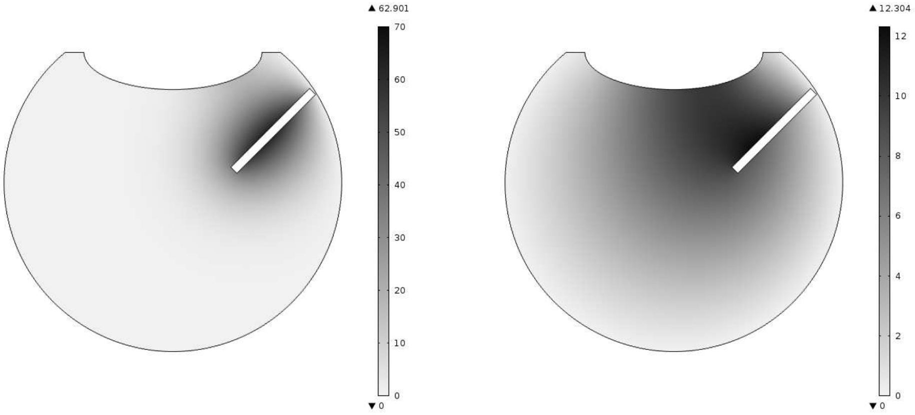


Figure 6: Drug concentration in the vitreous chamber at $5\,min$ (left) and $1\,h$ (right).

During the first instants of the delivery process, no drug is observed in the vitreous, except near the ocular implant, and as time increases more drug concentration is available to diffuse. For a better understanding of the qualitative behaviour of the drug concentration in the vitreous chamber, we present in Figure 7, the plot of drug concentration *vs* time at a point located inside the vitreous chamber and close to the lens. It can be observed that the drug concentration increases until it attains a maximum value, at $t = 30\,min$; for $t > 30\,min$ the drug concentration decreases until no drug concentration is present in the ocular implant. This qualitative behaviour is in agreement with medical data, establishing that for a duration of T units of time the maximum concentration of drug is attained for $\overline{T}$, where $\frac{T}{4} < \overline{T} < \frac{T}{3}$.

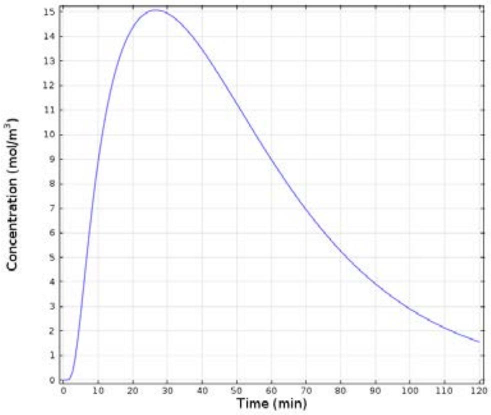


Figure 7: Drug concentration in the vitreous chamber along two hours.

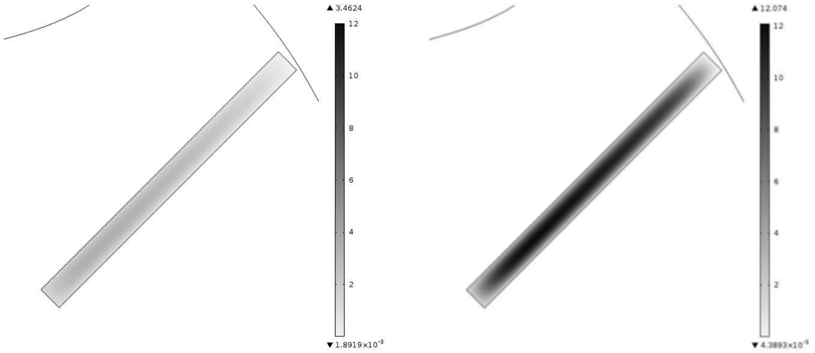


Figure 8: Drug concentration in the implant at $t = 2h$ - influence of degradation rate $\beta_1 = 5 \times 10^{-4}$ (left) and $\beta_1 = 1 \times 10^{-5}$ (right).

In Figure 8 the influence of the degradation rate is illustrated: a smaller value of β_1 leads to a slower degradation process and consequently more concentration is observed inside the polymeric implant.

In Figure 9 the influence of Young modulus is illustrated. As expected the increase of Young, E, delays the drug release and consequently more drug concentration is observed inside the polymer. In fact as crosslinking density is proportional to E, the large is this parameter, the more stiff is the material and a more significant barrier difficults the release of drug.

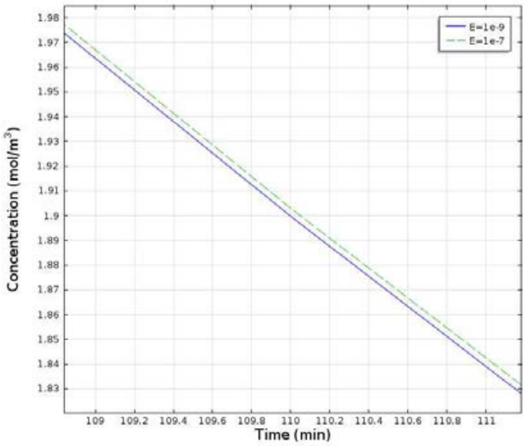


Figure 9: Drug concentration at a point of the boundary of the implant around $t = 110\,min$ - influence of E, $E = 1 \times 10^{-7}$ (top line) and $E = 1 \times 10^{-9}$ (down line).

5 Conclusion

A coupled model to simulate *in vivo* drug delivery from an intravitreal viscoelastic biodegradable implant has been developed. The whole process is described by a set of partial differential equations that take into account passive diffusion, convection resulting from the permeation of aqueous humor, stress driven diffusion and the degradation of the polymer. At the best of our knowledge the dynamics of drug desorption has not been described so far in the literature considering the simultaneous interplay between mechanical, physical and chemical effects. The numerical simulations show qualitative agreement with the physical expected behavior. The model clarifies the large influence of the degradation parameter in sustained drug delivery. The viscoelastic properties of the polymeric implant are also shown to be an effective control mechanism to delay or to speed up the release of drug. Mathematical modeling is a unique tool to explain transport mechanisms, and to help in implant design, avoiding expensive and extensive experimentation.

In future work physical values for all the parameters of the model should be retrieved. Also more realistic mechanical models will be considered and the heterogeneous structure of the vitreous, that is characteristic of elderly patients, should be taken into account.

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 projects PEst-C/MAT/UI0324/2011, SFR/BD/33703/2009, and by the project UTAustin/MAT/0066/2008.

References

[1] R. Avatar, D. Tandon, *A mathematical analysis of intravitreal drug transport*, Journal of Pharmaceutical Research, **7(1)**, (2008) 867-877.

[2] R. K. Balachandran and V. H. Barocas, *Computer modeling of drug delivery to the posterior eye: effect of active transport and loss to choroidal blood flow*, Pharmaceutical Research, **25** (2008) 2685-2696.

[3] R. K. Balachandran and V. H. Barocas, *Finite element modeling of drug distribution in the vitreous humor of the rabbit eye*, Annals of Biomedical Engineering **25** (1997) 303-314.

[4] C.Hal F. Brinson, L. Catherine Brinson, *Polymer Engineering Science and Viscoelasticity: An Introduction*, Springer, 2008.

[5] D. Cox, R. A. J. Phipps, B. Levine, A. Jacobs and D. Fowler, *Distribution of phencyclidine into vitreous humor*, Journal of Analytical Toxicology, **31**, (2007) 537-539.

[6] N. Haghjou, M. J. Abdekhodaie, Y. L. Cheng. M. Saadatmand, *Computer modeling of drug distribution after intravitreal administration*, W.A.S. Engineering and Tecnology, **53**, (2011) 706-716.

[7] J. Kathawate and S. Acharya, *Computational modeling of intravitreal drug delivery in the vitreous chamber with different vitreous substitutes*,International Journal of Heat and Mass Transfer,**51**, (2008) 5598-5609.

[8] C. W. Misner, K. S. Thorne and J. A. Wheeler, *Computer modeling of drug delivery to the posterior eye: effect of active transport and loss to choroidal blood flow*, Freeman, San Francisco, 1970.

[9] M. S. Stay, J. Xu, T. W. Randolph, and V. H. Barocas, *Computer simulation of convective and diffusive transport of controlled-release drug in the vitreous humor*, Pharmaceutical Research, **20(1)**, (2003) 96-102.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Analytical and Numerical Study of Diffusion through Biodegradable Viscoelastic Materials

E. Azhdari[1], J. A. Ferreira[1], P. de Oliveira[1] and P. M. da Silva[2]

[1] *CMUC, Department of Mathematics, University of Coimbra, 3001-454, Coimbra, Portugal*

[2] *Department of Physics and Mathematics, ISEC, Rua Pedro Nunes, Quinta da Nora, 3030-199 Coimbra, Portugal*

emails: ebrahim@mat.uc.pt, ferreira@mat.uc.pt, poliveir@mat.uc.pt, pascals@isec.pt

Abstract

In this paper the transport of a drug through a viscoelastic biodegradable material is studied. The phenomenon is described by a set of three coupled partial differential equations that take into account passive diffusion, stress driven diffusion and the degradation of the material. The stability properties of the system are studied. Numerical simulations show an influence of viscoelastic and degradation parameters in agreement with the expected physical behaviour.

Key words: viscoelasticity, Young modulus, degradation

1 Introduction

In the past few decades biodegradable polymers have attracted the attention of many researchers mainly for their applications in controlled drug delivery [5]. In this paper we will consider transport of a drug through a viscoelastic and hydrolyzed polymeric matrix.

The main actors in drug delivery are the living system, the composition of drug, the polymeric matrix where it is dispersed and the external conditions of release as for example the presence of an electric field or a heat source. To obtain a predefined release profile the mechanisms of control can act essentially on the polymeric matrix and the external conditions. In this paper we study a mathematical model to predict the influence of the mechanical and chemical properties of the polymer - viscoelasticity and degradation- in the release rate. As degradation proceeds, the polymer molecular weight decreases and diffusional paths open through the matrix allowing solved drug molecules to leave the device [6]. Because of the increasing permeability of the system upon polymer degradation, the constant diffusion coefficient is replaced by a molecular weight dependent diffusion coefficient [7]. The viscoelastic behaviour of the polymeric matrix is described by a Maxwell fluid model [1, 2, 3].

The paper is organised as follows. In Section 2 the mathematical model is presented. In Section 3 the qualitative behaviour of the released mass is studied. A fully discrete method that mimics the properties of the continuous problem is described in Section 4 and numerical simulations are exhibited in Section 5. Finally in Section 6 some conclusions are addressed.

2 The mathematical model

We consider a polymer filling a bounded domain $\Omega \subseteq \mathbb{R}^n$ with boundary $\partial\Omega$. The diffusion of drug from this polymer is described by the following system of partial differential equations:

$$\begin{cases} \dfrac{\partial C}{\partial t} = \nabla(D(M)\nabla C) + \nabla(D_v \nabla\sigma) \; \texttt{in} \; \Omega \times (0,T], \\[2ex] \dfrac{\partial \sigma}{\partial t} + \dfrac{E}{\mu}\sigma = EC \; \texttt{in} \; \Omega \times (0,T], \\[2ex] \dfrac{\partial M}{\partial t} + \beta_1 M = \beta_2 C \; \texttt{in} \; \Omega \times (0,T], \end{cases} \tag{1}$$

where C represents the unknown concentration of the drug inside the polymer, σ is the unknown stress, M is the unknown molecular weight of the polymer. The viscoelastic influence in the drug transport is represented by the term $\nabla(D_v \nabla\sigma)$ where D_v is the so called viscoelastic diffusion coefficient. This term states that the polymer acts as a barrier to the diffusion: as the drug strains the polymer it reacts with a stress of opposite sign. To account for the increasing permeability of the system upon polymer degradation, the diffusion coefficient is defined as

$$D(M) = D_0 e^{\frac{M_0}{M+M_0}},$$

where D_0 is the diffusion coefficient of drug in the non hydrolyzed polymer and M_0 is its initial molecular weight. The second equation in (1) defines the viscoelastic behaviour of the polymer by the Maxwell fluid model [2, 3]

$$\frac{\partial \sigma}{\partial t} + \frac{E}{\mu}\sigma = E\frac{\partial \epsilon}{\partial t}$$

where E represents the Young modulus of the material, μ its viscosity and ϵ is the strain produced by the drug molecules. If we assume that

$$\epsilon = k \int_0^t C(x,s)ds$$

where k is a positive constant we obtain the second equation in (1) where this constant has been absorbed by E. In the third equation of (1) β_1 and β_2 are constants that characterize the degradation properties of the material.

System (1) is completed with initial conditions

$$\begin{cases} C(x,0) = C_0, \ x \in \Omega, \\ \sigma(x,0) = \sigma_0, \ x \in \Omega, \\ M(x,0) = M_0, \ x \in \Omega, \end{cases}$$

and boundary conditions

$$\begin{cases} C(x,t) = 0 \ \texttt{ on } \partial\Omega \times (0,T], \\ \sigma(x,t) = 0 \ \texttt{ on } \partial\Omega \times (0,T], \\ M(x,t) = 0 \ \texttt{ on } \partial\Omega \times (0,T], \end{cases}$$

where $\partial\Omega$ denotes the boundary of Ω.

3 Qualitative behaviour of a mass related functional

In this section we study the qualitative behaviour of the mass related functional

$$M(t) = \int_\Omega C^2(t)dx, \quad t \geq 0.$$

From the second equation of (1) we easily get

$$\sigma(t) = E \int_0^t e^{-\frac{E}{\mu}(t-s)} C(s)ds + \sigma(0)e^{-\frac{E}{\mu}t}, \quad t \geq 0.$$

Replacing in the first equation of (1) we obtain for C

$$\frac{\partial C}{\partial t} = \nabla(D(M)\nabla C) + E \int_0^t e^{-\frac{E}{\mu}(t-s)} \nabla(D_v \nabla C(s))ds \ \ in \ \ \Omega \times (0,T]. \qquad (2)$$

As $\frac{1}{2}M'(t)=\int_\Omega C(t)\frac{\partial C}{\partial t}(t)dx$ we deduce, considering (2)

$$\frac{1}{2}M'(t)=-\Big\|\sqrt{D(M)}\nabla C(t)\Big\|^2-\Big(E\int_0^t e^{-\frac{E}{\mu}(t-s)}D_v\nabla C(s)ds,\nabla C(t)\Big), \tag{3}$$

where $(.,.)$ stands for the scalar product in $L^2(\Omega)$. From (3) we have

$$\frac{1}{2}M'(t)+\overline{D}_0\Big\|\nabla C(t)\Big\|^2\leq\frac{D_v^2}{4\epsilon^2}E^2\Big\|\int_0^t e^{-\frac{E}{\mu}(t-s)}\nabla C(s)ds\Big\|^2+\epsilon^2\Big\|\nabla C(t)\Big\|^2,$$

where $\overline{D}_0\leq D$ and $\epsilon\neq 0$. Consequently we deduce

$$\frac{1}{2}M'(t)+(\overline{D}_0-\epsilon^2)\Big\|\nabla C(t)\Big\|^2\leq\frac{D_v^2}{4\epsilon^2}E^2\int_0^t e^{-2\frac{E}{\mu}(t-s)}ds\int_0^t\Big\|\nabla C(s)\Big\|^2ds,$$

and then

$$M(t)+2(\overline{D}_0-\epsilon^2)\int_0^t\Big\|\nabla C(s)\Big\|^2ds\leq\frac{D_v^2}{2\epsilon^2}\frac{E^2}{2\frac{E}{\mu}}\int_0^t\int_0^s\Big\|\nabla C(\mu)\Big\|^2d\mu ds+M(0).$$

If ϵ^2 is such that

$$\overline{D}_0-\epsilon^2>0$$

we obtain

$$\begin{aligned}M(t)+\int_0^t\Big\|\nabla C(s)\Big\|^2ds &\leq \frac{D_v^2E^2}{\max\{1,2(\overline{D}_0-\epsilon^2)\}4\epsilon^2\frac{E}{\mu}}\int_0^t\int_0^s\Big\|\nabla C(\mu)\Big\|^2d\mu ds\\ &+ \frac{1}{\max\{1,2(\overline{D}_0-\epsilon^2)\}}M(0).\end{aligned}$$

Finally Gronwall's Lemma [4] leads to

$$M(t)+\int_0^t\Big\|\nabla C(s)\Big\|^2ds\leq\frac{1}{\max\{1,2(\overline{D}_0-\epsilon^2)\}}M(0)e^{\frac{D_v^2E^2}{\max\{1,2(\overline{D}_0-\epsilon^2)\}4\epsilon^2\frac{E}{\mu}}t}. \tag{4}$$

This last inequality establishes that $M(t)$ and $\int_0^t\|\nabla C(s)\|^2ds$ are bounded for bounded intervals of time. We note that (4) can be improved by eliminating the exponential factor in its right hand side [8]. A stability result of type

$$M(t)+\int_0^t e^{-2\gamma(t-s)}\Big\|\nabla C(s)\Big\|^2ds\leq M(0)$$

can then be stated for a convenient choice of the parameters of the model.

4 A discrete model

In order to simplify the presentation we consider in what follows $\Omega = (0,1)$. We fix $h > 0$ and we introduce in $\overline{\Omega}$ the grid

$$I_h = \{x_i, i = 0, \dots, N, x_0 = 0, x_N = 1, x_i - x_{i-1} = h, i = 1, \dots, N\}.$$

Discretizing the spatial derivative using the second order finite difference discretization

$$\frac{\partial}{\partial x}(D(M)\frac{\partial C}{\partial x})(x_i,t) \simeq \frac{D(\frac{M(x_i,t)+M(x_{i+1},t)}{2})D_{-x}C(x_{i+1},t) - D(\frac{M(x_i,t)+M(x_{i-1},t)}{2})D_{-x}C(x_i,t)}{h},$$

where D_{-x} represents the backward finite difference operator. We replace (1) by the following ordinary differential system

$$\begin{cases} \dfrac{dC_i(t)}{dt} = \dfrac{1}{h}\left(D(A_hM_i(t))D_{-x}C_{i+1}(t) - D(A_hM_{i-1}(t))D_{-x}C_i(t)\right) + D_vD_{2,h}\sigma_i(t), \\[2ex] \dfrac{d\sigma_i(t)}{dt} + \dfrac{E}{\mu}\sigma_i(t) = EC_i(t), \\[2ex] \dfrac{dM_i(t)}{dt} + \beta_1 M_i(t) = \beta_2 C_i(t), \end{cases} \tag{5}$$

and where, for $i = 1, \dots, N-1$, $C_i(t)$, $\sigma_i(t)$ and $M_i(t)$ stand for semi-discrete approximation of $C(t)$, $\sigma(t)$ and $M(t)$, respectively. In (5) A_h represents the average operator

$$A_h v(x_i) = \frac{1}{2}(v(x_i) + v(x_{i+1})),$$

and $D_{2,h}$ is the second-order finite difference operator

$$D_{2,h}u(x_i) = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}, \quad i = 1, \dots, N-1.$$

To solve system (5) we use the discretized boundary conditions

$$C_0(t) = C_N(t) = \sigma_0(t) = \sigma_N(t) = M_0(t) = M_N(t) = 0,$$

and the initial conditions

$$C_i(0) = C_0, \sigma_i(0) = \sigma_0, M_i(0) = M_0, \quad i = 1, \dots, N.$$

The time integration of (5) is performed considering as implicit-explicit approach, defined by

$$\begin{cases} \dfrac{C_i^{n+1}-C_i^n}{\Delta t} = \dfrac{1}{h}\left(D(A_hM_i^n)D_{-x}C_{i+1}^{n+1} - D(A_hM_{i-1}^n)D_{-x}C_i^{n+1}\right) + D_vD_{2,h}\sigma_i^n, \\ \dfrac{\sigma_i^{n+1}-\sigma_i^n}{\Delta t} + \dfrac{E}{\mu}\sigma_i^n = EC_i^{n+1}, \\ \dfrac{M_i^{n+1}-M_i^n}{\Delta t} + \beta_1 M_i^n = \beta_2 C_i^{n+1}. \end{cases} \tag{6}$$

In (6) C_i^n, σ_i^n, and M_i^n for $i = 1, ..., N$ stand for time for approximations of $C_i(t_n)$, $\sigma_i(t_n)$ and $M_i(t_n)$ in the time grid defined by

$$\{t_n, n = 0, \dots, M, t_0 = 0, t_M = T, t_n - t_{n-1} = \Delta t, n = 1, \dots, M-1\}.$$

System (6) is completed with the following conditions

$$C_0^n = C_N^n = \sigma_0^n = \sigma_N^n = M_0^n = M_N^n = 0, \quad n = 0, \dots, M.$$

It can be shown that method (6) is second order consistent in space and first order in time. In the numerical simulations, that we present in Section 5, we consider the drug released mass defined by

$$M(t) = \int_\Omega C(x,0)dx - \int_\Omega C(x,t)dx, \tag{7}$$

for each $t \in [0,T]$.

5 Numerical results

In this section we illustrate the use of the numerical scheme (6). We take $C_0 = 1, M_0 = 0.5, \sigma_0 = 0.5, D_0 = 0.01, D_v = -1 \times 10^{-4}, \mu = 1 \times 10^{-2}, E = 1 \times 10^{-3}, \beta_1 = 0.1, \beta_2 = 1 \times 10^{-3}$, $\Delta t = 5 \times 10^{-4}, h = 1 \times 10^{-2}$. The units of the concentration are mol/mm^3. The units of other variables and parameters are such that the equations are dimensionally correct.

In Figure 1 the evolution of C in time is illustrated. As expected the drug concentration decreases in time. The evolution of M is plotted in Figure 2, where the decrease in time of the molecular weight is consequence of the polymer degradation. In order to see that the evolution inside the polymer is not spatially homogeneous, a plot of the molecular weight at $t = 4$ is presented in Figure 2(right).

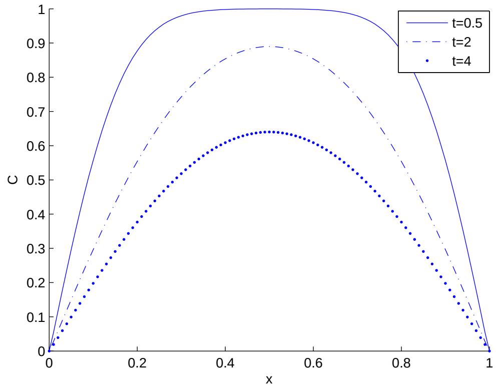


Figure 1: Concentration at different times.

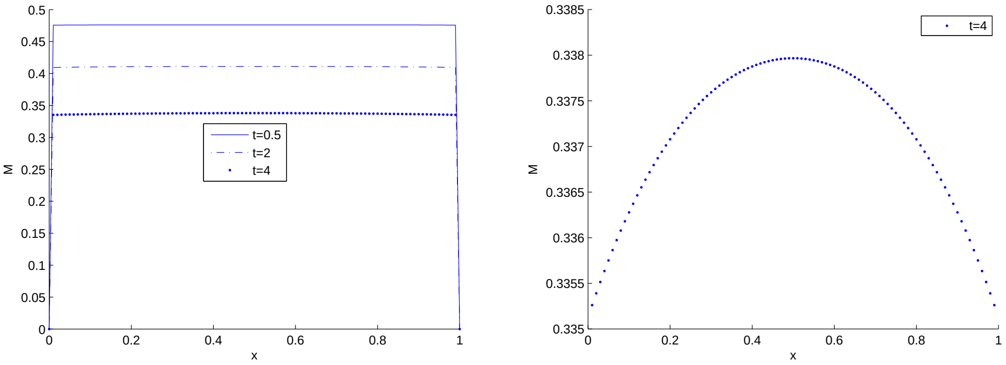


Figure 2: Molecular weight (left) at different times and a zoom of molecular weight at $t = 4$ (right).

In Figure 3 the influence of diffusion on released mass and molecular weight are shown for $t = 0.5,\ 2,\ 4$. As D_0 increases the released mass increases because the diffusion process becomes faster. Consequently as D_0 increases the concentration inside the polymer decreases and from the third equation in (1) we conclude that the molecular weight decreases. Obviously that taking into account the non linear character of the problem this argument is naive and it can not be considered a general result. However in Figure 3 (right) we illustrate this ansatz for the data used in our simulations.

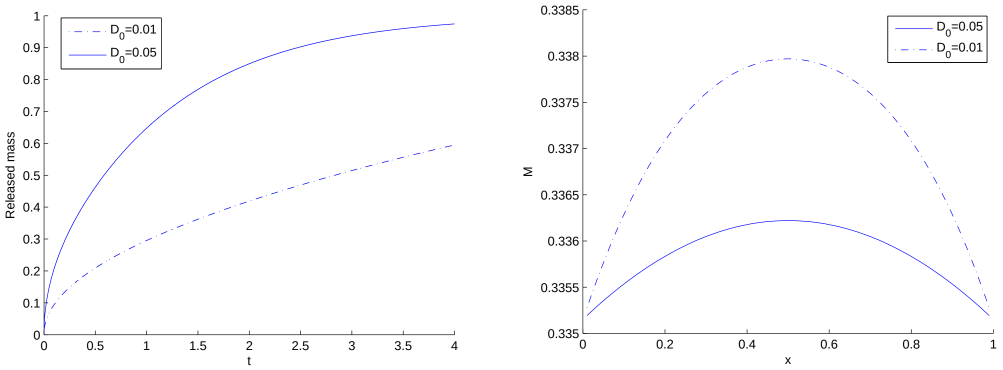


Figure 3: Influence of the diffusion on the released mass (left) and the molecular weight (right).

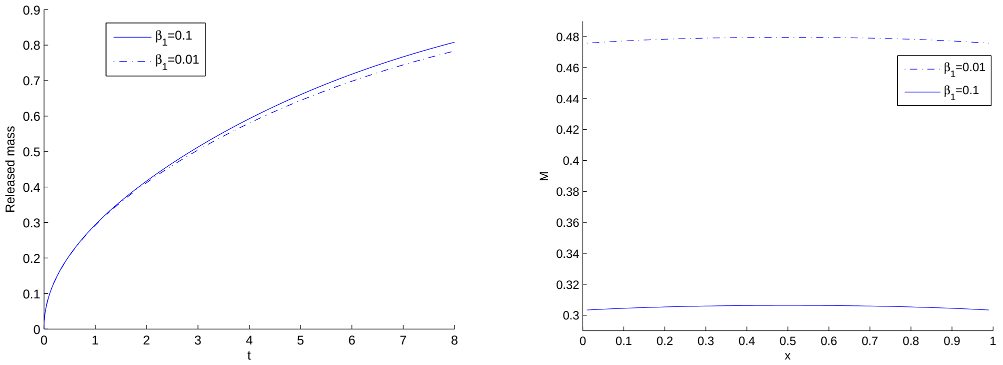


Figure 4: Influence of the degradation rate on the released mass (left) and molecular weight at $t = 8$ (right).

The influence of the degradation rate is presented in Figure 4. As expected if the degradation rate increases the rate delivery of the drug also increases. In the right of Figure

4 we observe that the increase of the degradation rate is closely related with loss of molecular weight.

In Figure 5 we study the dependance of released mass on the viscoelastic diffusion coefficient D_v. We observe that the polymer acts as a barrier that difficults drug diffusion. The drug molecules strain the polymer and it exerts a stress of opposite sign. The non Fickian flux $-D_v(\nabla\sigma)$ is, in a certain sense a antiflux which decreases the Fickian flux $-D(\nabla C)$. From a mathematical point of view we represent this interpretation by considering $D_v < 0$. In agreement with this description the increase of $\mid D_v \mid$ leads to a delay of release.

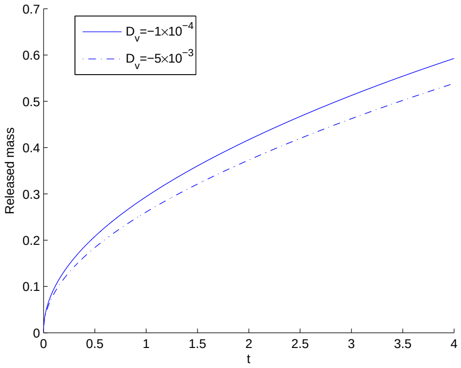


Figure 5: Influence of parameter D_v on the released mass.

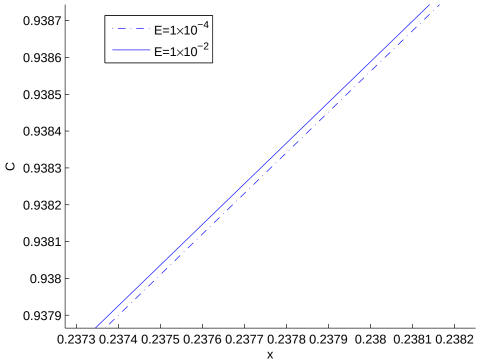


Figure 6: Influence of parameter E in the drug concentration at $t = 0.5$.

In Figure 6 the influence of Young modulus, E, in the drug concentration inside the polymer is presented at $t = 0.5$. The crosslink density of the polymer is proportional to

Young modulus E and consequently as this constant increases the polymer offers more resistance to the exit of the drug, which is delayed.

6 Conclusion

A model to simulate transport through a biodegradable viscoelastic material is studied. The analytical treatment of the system of partial differential equations lead to the establishment of stability results. The influence of mechanical and degradation parameters is analysed, showing agreement with physical behaviour. We believe that with future improvements the model can be used as a tool to design biodegradable polymers with predefined properties.

Acknowledgements

This work was partially supported by the Centro de Matemática da Universidade de Coimbra (CMUC) and by the Portuguese Government through the FCT - Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0324/2011, and Fundação para a Ciência e a Tecnologia, Grant SFRH / BD / 33812 / 2009.

References

[1] H. F. Brinson, L. C. Brinson, *Polymer engineering science and viscoelasticity, An introduction*, Springer, 2008.

[2] D. S. Cohen, A. B. White, JR. t, and T. P. Witelski, *Shock Formation in a Multidimensional Viscoelastic Diffusive System*, SIAM J. APPL. MATH **55(2)** (1995) 348–368.

[3] D. S. Cohen, A. B. White, JR, *Sharp Fronts due to Diffusion and Viscoelastic Relaxation in Polymers*, SIAM J. APPL. MATH **51(2)** (1991) 472–483.

[4] T. H. Gronwall, *Note on the derivative with respect to a parameter of the solutions of a system of differential equations*, Ann. of Math, **20(4)** (1919) 292-296.

[5] L. L. Lao, S. S. Venkatraman, N. A. Peppas, *Modeling of drug release from biodegradable polymer blends*, European Journal of Pharmaceutics and Biopharmaceutics, **70** (2008) 796–803.

[6] G. Perale, P. Arosio, D. Moscatelli, V. Barri, M. Müller, S. Maccagnan and M. Masi, *Anew model of resorbable device degradation and drug release: Transient 1-dimension diffusional model*, Journal of Controlled Release **136** (2009) 196–205.

[7] J. Siepmann, A. Göpferich, *Mathematical modeling of bioerodible, polymeric drug delivery systems*, Advanced Drug Delivery Reviews **48** (2001) 229–247.

[8] V. Thomée and L. B. Wahlbin, *Long-Time Numerical Solution of a Parabolic Equation with Memory*, Mathematics of Computation **62** (1994) 477–496.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Boolean sum 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

A boolean sum differential quadrature method is proposed. It combines an interpolation operator having a fundamental function with minimal compact support and a quasi-interpolation operator exact on the space of polynomials reproduced by the first one. We recall the resulting numerical scheme in the quintic case and derive new schemes from low degree B-splines.

Key words: differential quadrature, B-spline, interpolation, discrete quasi-interpolation

1 Introduction

The Differential Quadrature Method (DQM) is a numerical discretization technique for the approximation of derivatives by means of weighted sums of function values. It was introduced by Bellman and coworkers in the early 1970's, and it has been extensively employed to approximate spatial partial derivatives (cf. [2], [8] and references quoted therein). The classical DQM is polynomial-based, but some spline based DQMs have been proposed the limitation concerning the number of grid points involved. Given a B-spline (cf. [3], [9]), a cardinal lagrangian o hermitian spline with a compactly supported fundamental function is defined, from which the approximation of the derivatives is derived. The construction of these spline interpolants depends strongly on the degree of the B-spline (see for instance [4] and [10]). In this work we present a DQM based on interpolation and quasi-interpolation. Firstly, we consider the construction of compactly supported cardinal functions L based on B-splines such that

$$L(j) = \delta_{j,0},\ j \in \mathbb{Z}, \tag{1}$$

δ being the Kronecker sequence. Then, we revise some spline discrete quasi-interpolants defined from the same B-splines (cf. e.g. [3], [9] and references quoted therein). Finally, both the interpolants and the quasi-interpolants are used to define new interpolants having compactly supported fundamental functions again, and the maximal order of approximation. The quintic case is described and compared with the results obtained in [10].

2 Spline interpolation

Let M_n be the B-spline of order $n \geq 2$ centered at the origin. We look for a symmetrical fundamental function L such that

$$L = \sum_{j\in J} c_j M_n\left(2\cdot -j\right) \tag{2}$$

for some $c_j \in \mathbb{R}$. Once the structure for L is fixed, we prove that there exist such functions satisfying the interpolation conditions (1) and we determine one of them.

We will use the following notations. For each $r \geq 0$, $e_r\left(z\right) := z^r$ is the monomial of degree r, $\mathbb{P}_r$ denotes the space of polynomials of degree at most r, and $\widetilde{M_n} := \sum_{j\in\mathbb{Z}} M_n\left(j\right) e_j$ stands for the symbol of M_n.

Lemma 1 *The Laurent polynomials*

$$\Phi_0 := \sum_{j\in\mathbb{Z}} M_n\left(2j\right) e_{2j} \quad \textit{and} \quad \Phi_1 := \sum_{j\in\mathbb{Z}} M_n\left(2j+1\right) e_{2j+1}$$

have no common zeros on $\mathbb{C}\setminus\{0\}$.

As a consequence of the previous lemma (see [7]), 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. We use this result in order to obtain a symmetric fundamental function with a small support.

Proposition 2 *For each* $n \geq 4$, *let* $J := \{-d_n, \ldots, d_n\}$ *where*

$$d_n := \begin{cases} \lfloor r \rfloor - 2, & \textit{for } n \textit{ even}, \\ \lfloor r \rfloor - 1, & \textit{for } n \textit{ odd}, \end{cases}$$

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*

$$L = a_0 M_n\left(2\cdot +d_n\right) + \cdots + a_{2d_n} M_n\left(2\cdot -d_n\right)$$

satisfies conditions (1). It follows that

$$\operatorname{supp} L \subset \begin{cases} \left[-\frac{n}{2}+1, \frac{n}{2}-1\right], & \textit{for } n \textit{ even,} \\ \left[-\frac{n}{2}+\frac{3}{4}, \frac{n}{2}-\frac{3}{4}\right], & \textit{for } n \textit{ odd.} \end{cases}$$

For a given function f defined on the real line, the spline

$$\mathcal{L}(f) = \sum_{i\in\mathbb{Z}} f(i)\, L(\cdot - i)$$

interpolates f at the integers. Its associated scaled operator

$$\mathcal{L}_h(f) = \sum_{i\in\mathbb{Z}} f(hi)\, L\left(\frac{\cdot}{h} - i\right)$$

permit us to interpolate f at $h\mathbb{Z}$, $h > 0$.

3 Discrete quasi-interpolation

In practice we need a spline interpolant $\mathcal{I}(f)$ such that its associated scaled operator $\mathcal{I}_h$ provides the maximal approximation order, i.e.

$$\left\|\mathcal{I}_h(f) - f\right\|_\infty \leq Ch^n \left\|f^{(n)}\right\|_\infty$$

for some constant C independent of f and h, when $f \in C^n(\mathbb{R})$ with $f^{(n)} \in L^\infty(\mathbb{R})$. To construct $\mathcal{I}$, we will consider an appropriate discrete quasi-interpolant based on the B-spline M_n. A such discrete quasi-interpolant $\mathcal{Q}f$ for a given function f is a linear combination of integer translates of M_n. It can be written as

$$\mathcal{Q}f = \sum_{i\in\mathbb{Z}} f(i)\, q_n(\cdot - i), \tag{3}$$

with $q_n = \sum_{j=-m}^{m} \gamma_j M_n(\cdot - j)$. In general, the compactly supported function q_n does not satisfy conditions (1). There are different methods to determine the coefficients γ_j but the exactness of $\mathcal{Q}$ on $\mathbb{P}_{n-1}$ is required, i.e. $\mathcal{Q}(p) = p$ for all $p \in \mathbb{P}_{n-1}$.

Let $\mathcal{Q}$ be a quasi-interpolant exact on $\mathbb{P}_{n-1}$. The order of approximation of $\mathcal{L}$ is increased by forming the boolean sum $\mathcal{L}\oplus\mathcal{Q}$ of $\mathcal{L}$ and $\mathcal{Q}$, defined by the expression $\mathcal{L}\oplus\mathcal{Q} = \mathcal{L}+\mathcal{Q} - \mathcal{L}\mathcal{Q}$. It is well known (cf. [5, Section 9.4]) that $\mathcal{L} \oplus \mathcal{Q}$ inherits the interpolation properties of $\mathcal{L}$, the exactness of $\mathcal{Q}$, and produces $C^{n-2}(\mathbb{R})$ functions.

4 Boolean sum based differential quadrature

We propose to define the interpolation operator $\mathcal{I}$ as the boolean sum of $\mathcal{L}$ and a quasi-interpolation operator, $\mathcal{Q}$, i.e. $\mathcal{I} = \mathcal{L} \oplus \mathcal{Q}$. From (2) and (3), we can write

$$\mathcal{I}(f) = \sum_{i\in\mathbb{Z}} f(i)\, L(\cdot - i) + \sum_{i\in\mathbb{Z}} \left(Q(f)(i) - f(i)\right) q(\cdot - i).$$

We use this expression to approximate the derivative of a function with respect to a space variable.

We recall the quintic case (see [1]) based on a Chebyshev quasi-interpolation operator and compare this two-stage method with the scheme proposed in [10] for a given function defined on the real line. In practice this function is defined on an interval, and standard additional modifications are required near the boundary.

In [10] the following fundamental function is constructed

$$\begin{aligned}\varphi = {} & \frac{20523}{440} M_5 - \frac{3483}{110}\left(M_5\left(\cdot + \frac{1}{3}\right) + M_5\left(\cdot - \frac{1}{3}\right)\right) \\ & + \frac{8829}{880}\left(M_5\left(\cdot + \frac{2}{3}\right) + M_5\left(\cdot - \frac{2}{3}\right)\right) - \frac{131}{110}\left(M_5(\cdot + 1) + M_5(\cdot - 1)\right).\end{aligned}$$

It is supported on $[-4, 4]$, and the corresponding quintic interpolation spline

$$\mathcal{Z}(f) = \sum_{i\in\mathbb{Z}} f(i)\, \varphi(\cdot - i) \tag{4}$$

is defined on the uniform partition $\frac{1}{3}\mathbb{Z}$ of $\mathbb{R}$.

It is easy to derive the expression

$$\mathcal{Z}(f)'(i) = -\frac{1}{60} f(i-3) + \frac{3}{20} f(i-2) - \frac{3}{4} f(i-1) + \frac{3}{4} f(i+1) - \frac{3}{20} f(i+2) + \frac{1}{60} f(i+3) \tag{5}$$

for the derivative of $\mathcal{Z}(f)$ at $i \in \mathbb{Z}$. Although $\mathcal{Z}$ is an operator exact on $\mathbb{P}_5$, the obtained formula is exact on $\mathbb{P}_6$.

We give the error of this formula for $f \in C^7(\mathbb{R})$.

Proposition 3 *Let $i \in \mathbb{Z}$. For any $f \in C^7(\mathbb{R})$, the error for the scheme given by (5) satisfies*

$$\left|f'(i) - Z(f)'(i)\right| \leq \frac{1}{140}\left\|f^{(7)}\right\|_{\infty,[i-3,i+3]}, \tag{6}$$

where $\|g\|_{\infty,I}$ stands for the uniform norm of g on the interval I.

The method proposed here is obtained by combining a quintic spline interpolation operator, $\mathcal{L}$, and a quintic quasi-interpolant, $\mathcal{Q}$ (see [6]). From the results in Section 2, we get the fundamental function

$$L = -\frac{15}{208} M_6\left(2\cdot+1\right) + \frac{15}{8} M_6\left(2\cdot\right) - \frac{15}{208} M_6\left(2\cdot-1\right).$$

It is a spline function defined on the uniform partition $\frac{1}{2}\mathbb{Z}$ of $\mathbb{R}$. It has support $[-2, 2]$.

We also consider the quasi-interpolant (see [6])

$$\mathcal{Q}f = \sum_{i\in\mathbb{Z}} \left(\sum_{j=-3}^{3} \gamma_j f\left(i-j\right) \right) M_6\left(\cdot - i\right), \tag{7}$$

where

$$\gamma = \left(-\frac{353}{30720}, \frac{1891}{15360}, -\frac{19631}{30720}, \frac{15781}{7680}, -\frac{19631}{30720}, \frac{1891}{15360}, -\frac{353}{30720} \right).$$

The quintic interpolation operator $\mathcal{I} = \mathcal{L} \oplus \mathcal{Q}$ produces the formula

$$\begin{aligned} \mathcal{I}\left(f\right)'\left(i\right) = {} & \frac{6}{38338560}\left(f\left(6\right) - f\left(-6\right)\right) - \frac{9}{26624}\left(f\left(5\right) - f\left(-5\right)\right) \\ & - \frac{10663}{9584640}\left(f\left(4\right) - f\left(-4\right)\right) + \frac{14341}{479232}\left(f\left(3\right) - f\left(-3\right)\right) \\ & - \frac{1422127}{7667712}\left(f\left(2\right) - f\left(-2\right)\right) + \frac{377275}{479232}\left(f\left(1\right) - f\left(-1\right)\right). \end{aligned} \tag{8}$$

It requires the values of f at the points in a larger subset than the Zhong's one. However, next result shows that the constant in the error expression is smaller than $\frac{1}{140} \approx 0.00714286$.

Proposition 4 *Let $i \in \mathbb{Z}$. For any $f \in C^7\left(\mathbb{R}\right)$, the error for the scheme given by (8) satisfies*

$$\left|f'\left(i\right) - I\left(f\right)'\left(i\right)\right| \le 0.00324516 \left\| f^{(7)} \right\|_{\infty,[i-3,i+3]}. \tag{9}$$

Formula (8) is more complex than (5), but the constant in (9), despite not being optimal, is about the 45% of the one in (6). Thus, we can use equation (8) to solve numerically a problem with a bigger step h. Moreover, we take advantage in using (7) instead of (4) to interpolate the numerical solution of that problem because its associated partitions are $\frac{1}{2}\mathbb{Z}$ and $\frac{1}{3}\mathbb{Z}$, respectively.

We will consider a general family of discrete quasi-interpolants in order to derive new differential quadrature methods by using the boolean sum based method.

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, 58 June 2006.

[2] C. W. Bert and M. Malik, Differential quadrature method in computational mechanics: a review. *Appl. Rev.*, **49**, 1–27, 1996.

[3] C. de Boor, *A practical guide to splines.* Springer-Verlag, New York, 2001.

[4] Q. Guo, H. Zhong, Non-linear vibration analysis y a spline-based differential quadrature method. *Journal of Sound and Vibration*, **269**, 413–420, 2004.

[5] J. Hoshek and D. Lasser, *Fundamentals of Computer Aided Geometric Design.* A. K. Peters, Wellesley, Massachussets, 1993.

[6] M. J. Ibáñez-Pérez, Chebyshev-type discrete quasi-interpolants, Mathematics and Computers in Simulation 77 (2008) 218–227.

[7] C. A. Micchelli, Banded matrices with banded inverses. *J. Comput. Appl. Math.*, **41**, 281–300, 1992.

[8] C. Shu, *Differential Quadrature and its applications in Engineering.* Springer-Verlag, London, 2000.

[9] L. L. Schumaker, *Spline functions. Basic theory.* John Wiley & Sons, New York, 1981.

[10] H. Zhong, Spline-based differential quadrature for fourth order differential equations and its applications to Kirchhoff plates. *Applied Mathematical Modelling*, **28**, 353–366, 2004.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

A WENO-based method to improve the determination of the threshold voltage and characterize DIBL effects in MOSFET transistors

D. Barrera[1], P. González[1], M. J. Ibáñez [1], A. M. Roldán [2] and J. B. Roldán [2]

[1] *Department of Applied Mathematics, Campus de Fuentenueva s/n, 18071-Granada, Spain, University of Granada*

[2] *Department of Electronics, Campus de Fuentenueva s/n, 18071-Granada, Spain, University of Granada*

emails: {dbarrera, prodelas, mibanez, amroldan, jroldan}@ugr.es

Abstract

A numerical procedure to obtain the threshold voltage in MOSFETs transistors has been developed by means of a weighted ENO scheme. Making use of this technique, the numerical noise that comes up in the experimental and simulated data usually employed to characterize these transistors is much reduced. The need of an accurate determination of these parameters motivates the use of this advanced numerical approach, that solves many of the issues that affect the conventional parameter extraction procedures currently in use in the microelectronics industry. In addition, also the influence of DIBL effects on the threshold voltage in short channel MOSFETs has been analyzed with this weighted ENO procedure.

Key words: MOSFETs, Threshold voltage extration, DIBL effects, WENO procedure
MSC 2000: AMS codes (65Z05, 65D05, 65D10)

1 Introduction

MOSFETs (Metal Oxide Semiconductor Field Effect Transistors) are the cornerstone in the current semiconductor industry since they are the most used devices for integrated circuit fabrication. Consequently, great efforts have been devoted in the last twenty years to characterize, simulate and model these MOSFET transistors. The scaling race, the reduction of transistor dimensions to improve the integrated circuit performance, has focused the research and development efforts in this industry since its very beginning. Because of this, billions of devices can be integrated in state-of-the-art chips [9], [12].

After the fabrication of these transistors, they undergone a characterization process, where the main electrical features: electric currents, capacitances and other magnitudes are measured versus the voltages applied at their terminals, and finally analyzed. These curves are used to determine relevant parameters that characterize the device operation. In particular, these parameters are introduced within compact models to help electronics engineers incorporate the new devices in the circuits they design.

From the compact modeling viewpoint, both the analytical expressions and the extracted parameters are important. The differences between several transistor technologies are included in a file of parameters that allow the inclusion of a certain fabrication technology in the engineers design tools. That is why an exact determination of these parameters is essential to reduce design costs. The extraction process is performed using numerical adjustments and regressions of hundreds of experimentally measured curves. In doing so, the approximation theory and the numerical techniques employed are essential, and in this context we present this paper.

In this work we will use a one-dimensional W.E.N.O. procedure [1], [10] to conveniently approximate the second derivative of a whole family of curves extracted from a two-dimensional drain current data set for MOSFET transistors. To do so, we will take advantage of the essentially non oscillatory nature of the corresponding polynomial WENO interpolation to extract the MOSFET transistor threshold voltage (V_T), which is known to be a key magnitude from the modeling viewpoint [8]. There are many different numerical techniques to determine the threshold voltage [4, 14] although we will focus in one of the most currently chosen in industry: the Transconductance Change Method (TCM) or Second Derivative Method (SDM) [3, 4, 5, 7, 13]. In the TCM method, the threshold voltage is related to the determination of the gate voltage at which the transistor low drain voltage derivative of the transconductance is maximum (the transconductance is obtained as the drain current derivative with respect to the gate voltage). Therefore, see Figure 1, this parameter is calculated by obtaining the maximum of the second derivative of the drain current versus gate voltage curve measured at constant low drain voltage (voltages are measured with respect to a grounded terminal: the source; V_{ds} and V_{gs} are equivalent to V_d and V_g respectively):

$$\left.\frac{\partial^2 I_d}{\partial V_g^2}\right|_{\text{measured at low } V_d, V_g=V_T} \quad \text{is a maximum.}$$

Making use of this development, we pretend to study the dependence of the threshold voltage on the drain voltage (V_d). This effect is essential in current transistors due to their reduced size; it is known to be produced by Drain Induce Barrier Lowering (DIBL) effects. In general, the higher the V_d the lower the V_T. It is well known in the literature that this

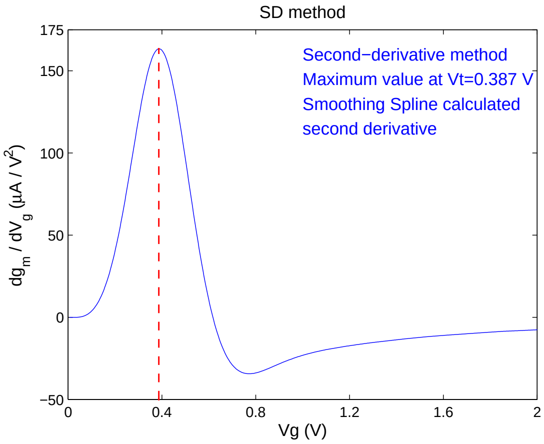


Figure 1: TCM or SDM using $I_d - V_g$ characteristics. In this method we seek the maximum of the second derivative of the $I_d - V_g$ curve. This value is taken as the threshold voltage of the device.

dependence can be approximately described with a linear relation [3].

In order to deepen on this issue, we have simulated several transfer characteristics $(I_d - V_g)$ for different drain voltages. We applied on this 2D data set of drain current values our technique to study the dependence commented before. We have analyzed if the advantages of the technique for 1D dataset (great reduction of numerical noise, good accuracy at calculating the second or higher order derivatives..., see ref. [1]) also hold when dealing with 2D data sets.

2 WENO methods

As it is well explained and documented in the literature, the WENO procedure (or Weighted ENO) is an improved technique, introduced by Liu et al. in [11], consisting in assigning to each subinterval all the possible stencils of a certain length containing it and constructing the interpolating polynomial as a particular convex combination of the corresponding polynomials. This is a better approach than in the ENO method, since now we will use all the information provided for the simpler ENO selection process, obtaining a higher order of accuracy at points in smooth regions of the function. The key then will be the appropriate assignments of weights to this convex combination in order to minimize the net contribution to the final combination of those polynomials corresponding to singularity-crossing stencils. In this way, the necessity of choosing appropriate “smoothness indicators” that will be capable of measuring adequately the smoothness of the function being interpolated is of vital

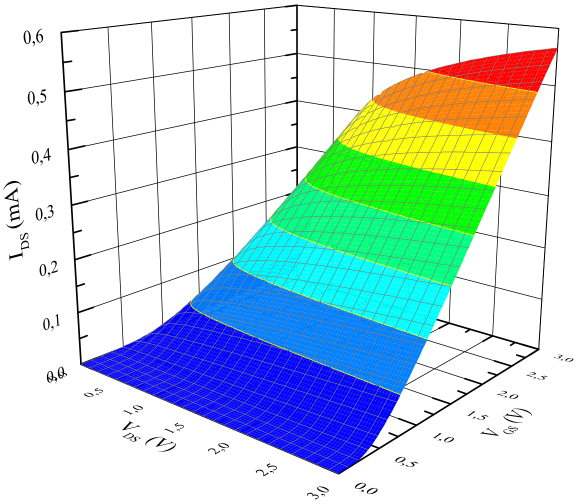


Figure 2: Drain current versus gate voltage and drain voltage for the transistor simulated.

importance. Jian and Shu [10] presented some of such indicators, usually more efficient than that proposed originally by Liu et al.

Another important issue of these techniques is that they were proposed originally for their application to the numerical resolution of hyperbolic problems and that they are formulated for the construction of the corresponding ENO or WENO interpolants, based mainly on cell averages and not just on point values. But some other authors, like those in [2] have adapted the original procedures to the corresponding point-value interpolation framework, that is what we also use here.

3 Transistor description

In order to analyze the numerical procedures for extracting the threshold voltage reported at the introduction, we have simulated a conventional silicon MOSFET transistor making use of ATLAS (a TCAD tool developed by Silvaco [6] for electronics device simulation, design and characterization). The main technological features of the transistor simulated are the following: $L = 0.3\,\mu$m channel doping concentration at the oxide surface level $N_A = 10^{17}\,\mathrm{cm}^{-3}$, and oxide thickness $T_{ox} = 8.7\,\mathrm{nm}$; a N$^+$ poly-silicon gate was used.

4 Results and discussion

The first results obtained with this WENO procedure are being very encouraging, and in good agreement with the results from the experimental measurements and other model

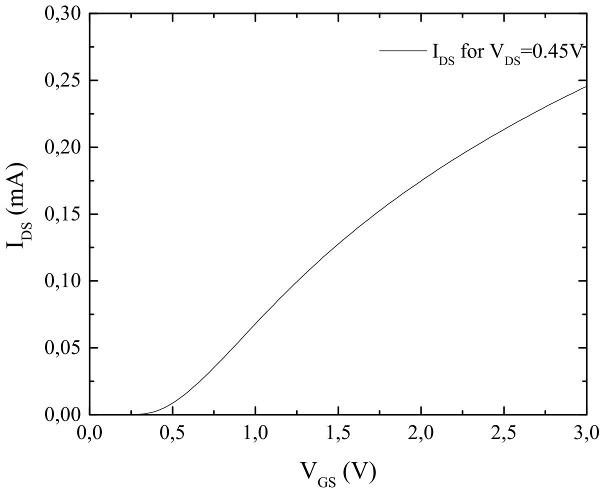


Figure 3: Drain current $I_d(A)$ versus gate voltage, $V_g(V)$, for a constant drain voltage V_d.

validations using different techniques, so they are making us thinking that they could be appropriate tools for many problems of this type.

In Figure 3, the simulated drain current versus gate voltage of the device under study is shown. This is one of the curves extracted from the data in Figure 2. The derivative of these data is shown in Figure 4. As it is well known, the numerical noise is much greater when calculating derivatives of the simulated and measured data, which is the case here. Nevertheless, the use of the WENO procedure, greatly reduces the oscillations. These effects are obviously greater for the second derivative; but this time, again, the use of an advance numerical procedure such as WENO produces a smooth data set (see Figure 5). This output allowed us to easily calculate the threshold voltage since the maximum is clearly defined and the typical errors connected with the numerical noise produced in calculating the derivatives are avoided.

The use of the previously described WENO procedure helped us with the analysis of the threshold voltage dependence on the drain voltage, due to the well known DIBL effects [3]. As can be seen, a linear trend shows up when the threshold voltage is calculated versus the drain voltage, as expected (see Figure 6). The use of this method presents promising possibilities for the automatic determination of the threshold voltage in different bias conditions using the TCM method. In this respect, the numerical stability linked to the WENO procedure could permit the inclusion of complex parameter extraction methods, numerically speaking, in conventional industrial tools.

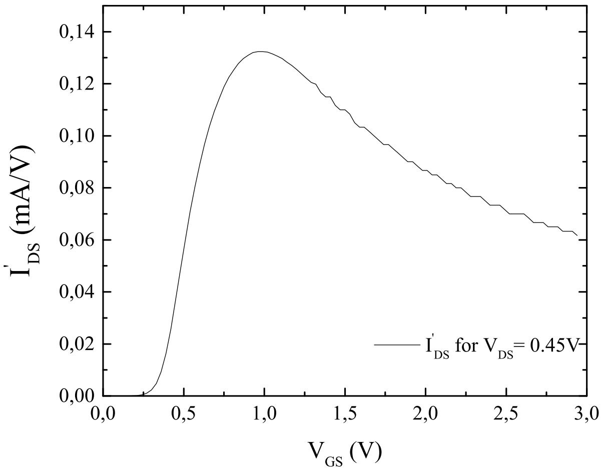


Figure 4: Approximation of the first derivative $I'_d(A/V)$ versus gate voltage, $V_g(V)$, using the WENO procedure.

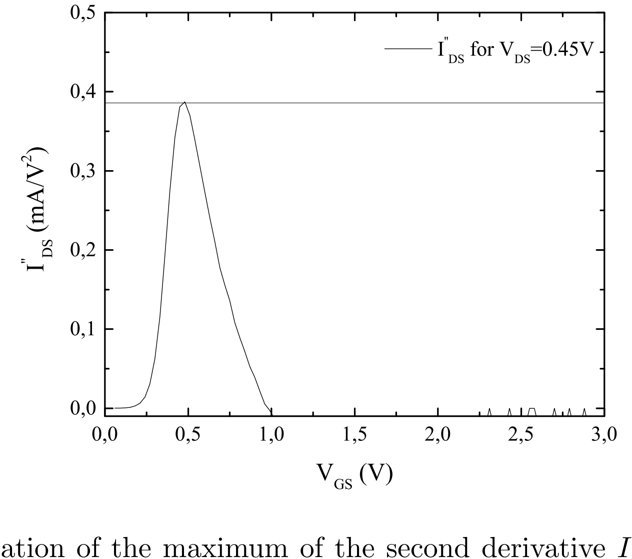


Figure 5: Approximation of the maximum of the second derivative $I''_d(A/V^2)$ versus gate voltage, $V_g(V)$.

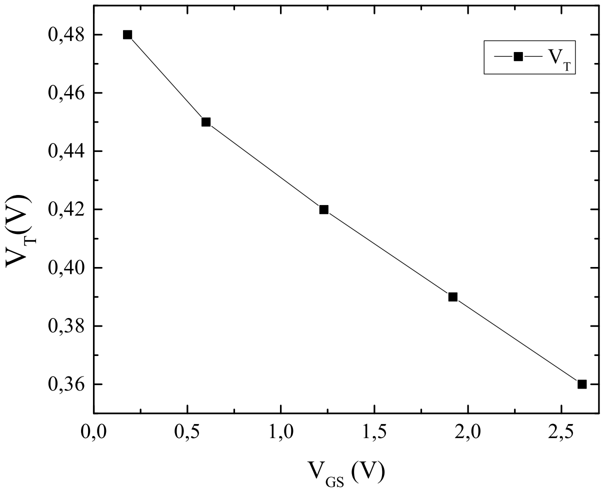


Figure 6: Calculated threshold voltage $V_T(V)$ versus drain voltage, $V_d(V)$, obtained by means of the transconductance change method, using the WENO approximations of the derivatives and a linear interpolation at specific points.

Acknowledgements

This work was partially carried out within the framework of Research Project P08-TIC-3580 and Resrarch Group FQM/191 supported by the Junta de Andalucía (Spain) and by the Spanish Ministerio de Ciencia e Innovación under grant MTM2011-26468.

References

[1] F. Arándiga and A. M. Belda, Weighted ENO interpolation and applications. Commnunications in Nonlinear Science and Numerical Simulation **9 (2)** 187–195 (2003).

[2] F. Arándiga, A. M. Belda and P. Mulet, Point-Value WENO Multiresolution Applications to Stable Image Compression. J. Sci. Computing **43** 158–182 (2010).

[3] N. Arora, *MOSFET modelling for VLSI simulation. Theory and practice*, reprinted by World Scientific, 2007.

[4] A. Ortiz-Conde, F.J. Garcia Sanchez, J.J. Liou, A. Cerdeira, M. Estrada, Y. YueN, *A review of recent MOSFET threshold voltage extraction methods*, Microelectronics Reliability 42, pp. 583–596, 2002.

[5] A. Ortiz-Conde, F.J. Garcia Sanchez, J. Muci, A. Teran, J. Liou, C. Ho, *Revisiting MOSFET threshold voltage extraction methods*, Microelectronics Reliability 53, 90–104, 2013.

[6] Atlas manuals. http://www.silvaco.com

[7] D. Flandre, V. Kilchytska and T. Rudenko, g_m/I_d method for threshold voltage extraction applicable in advanced MOSFETs with nonlinear behavior above threshold, IEEE Electron Device Letters 31 (9) (2010) 930–932.

[8] C. Galup-Montoro and Ch. Schneirer Marcio, *Mosfet modelling for circuit analysis and design*, World Scientific, 2007.

[9] The International Technology Roadmap for Semiconductors, ed. 2011 and 2012 update. http://public.itrs.net.

[10] G. Jiang and C.-W. Shu, Efficient implementation of weighted ENO schemes, Journal of Computational Physics **126** 202–228 (1996) from an ICASE-NASA Report of 1995.

[11] X-D. Liu, S. Osher and T. Chan, Weighted essentially non-oscillatory schemes, J. Comput. Phys., **115**, 200–212 (1994).

[12] G. Moore, Cramming more components onto integrated circuits, Electronics Magazine 38 (8) 1965.

[13] H. S. Wong, M. H. White, T. J. Kurttsick and R. V. Booth, Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFETs, Solid State Electronics 20 (1987) 953–968.

[14] M. J. Ibáñez, J. B. Roldán, A. M. Roldán, R. Yáñez, A comprehensive characterization of the threshold voltage extraction in MOSFETs transistors based on smoothing splines, Mathematics and Computers in Simulation, accepted for publication, (2013).

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Error estimates for modified Hermite interpolant on the simplex

D. Barrera[1], **M. J. Ibáñez**[1] **and O. Nouisser**[2]

[1] *Departamento de matemática aplicada, Universidad de Granada, Spain*

[2] *Département de mathématiques, Faculté Polydisciplinaire de Safi, Maroc*

emails: dbarrera@ugr.es, mibanez@ugr.es, otheman.nouisser@yahoo.fr

Abstract

The objective of this paper, is to study the error estimates of the Hermite interpolation operator of a multivariate function defined on a given regular simplex using the first and the second derivatives of the given function at the vertices of the simplex. The obtained operator is the identity of the cubic polynomials. When the data are given inside the simplex, we construct another approximation operator and we give the associated error estimates.

Key words: Interpolation, Hermite interpolation, Approximation order, error estiamtes.

1 Introduction

Let S be a simplex in the d-dimensional Eulidean space $\mathbb{R}^d$, with vertices $\nu_0, \cdots, \nu_d$. Suppose we know the values of a function f and those of its partial derivatives up to order two at the vertices of S. Our aim is to construct an approximation operator that interpolate f and its derivatives from this data.

Thomas F. Sturm [3] considered the Hermite interpolation operator

$$L[f](x) = \sum_{i=0}^{d} \left(f(\nu_i)\varphi_i(\lambda(x)) + \sum_{r=0, r\neq i}^{d} D_{\nu_r - \nu_i} f(\nu_i)\phi_{i,r}(\lambda(x)) \right)$$

that interpolate f and its first derivatives at the vertices of S. He proved that L reproduce the space of quadratic polynomials. The basis functions φ_i and ϕ_i $1 \leq i \leq d$, are the cubic polynomials given by

$$\begin{aligned}\varphi_i(\lambda(x)) &= \lambda_i\left(1+\lambda_i-\sum_{r=0}^{d}\lambda_r^2\right),\\ \phi(\lambda(x)) &= \frac{1}{2}\lambda_i\lambda_r\left(1+\lambda_i-\lambda_r\right)\end{aligned}$$

where $\lambda_i(x), 0 \leq i \leq d$, are the barycentric coordinates of x with respect to S that satisfy

$$\begin{aligned}x &= \sum_{i=0}^{d}\lambda_i(x)\nu_i,\\ \lambda_i(x) &\geq 0, \forall x \in S,\\ 1 &= \sum_{i=0}^{d}\lambda_i(x), \forall x \in S.\end{aligned}$$

Her, we denote by $D_y^j f(x)$ the jth directional derivative of f at x, defined by

$$D_y^j f(x) = \left[\frac{d^j}{dt^j}f(x+ty)\right]_{t=0}$$

In this paper and in order to improve the algebraic precision of the operator L, we consider a modified Hermite interpolation operator defined by :

$$H[f](x) = \sum_{i=0}^{d}\left\{\left(f(v_i)+\frac{1}{3}D_{x-v_i}f(v_i)\right)+\sum_{r=0,r\neq i}^{d}\left(\frac{2}{3}D_{v_r-v_i}f(v_i)+\frac{1}{3}D_{v_r-v_i}D_{x-v_i}f(v_i)\right)\phi_{i,r}(x)\right\}.$$

We establish an integral representation for the error of approximation considered by the operator H in terms of the one considered by L. In this setting, we establish the error estimates with respect the infinity norm. Finally, Numerical examples are illustrated to show the theoretical results.

2 representation of the error of the modified operator

By $C(\S)$, we denote the class of all real-valued continuous functions on S, and by $C^m(S)$, where $m \in \mathbb{N}$, the subclass of all functions that are m times continuously differentiable in the following sense. For each $x \in S$ and any $y \in \mathbb{R}^d$ such that $x + y \in S$, the directional derivatives $D_y^j f(x)$ exist for $j = 1, \cdots, m$ and depend continuously on x.
The error of the approximation by $H[f]$ can be represented as follows.

Théorème 2.1 *Let $f \in C^4(S)$. Then*

$$\begin{aligned} f(x) - H[f](x) &= \frac{1}{6}\sum_{i=0}^{d}\left(\int_0^1 D^4_{x-\nu_i} f(\nu_i + t(x-\nu_i))t^2(1-t)dt\right)\varphi_i(x) + \\ & \sum_{i=0}^{d}\sum_{r=0,r\neq i}^{d}\left(\int_0^1 D_{\nu_r-\nu_i}D^4_{x-\nu_i} f(\nu_i + t(x-\nu_i))dt\right)\phi_{i,r}(\lambda(x)) \end{aligned}$$

when the second derivatives are not involved, the operator H is reduced to the operator L, and the error of the approximation by $L[f]$ can be reprensented as follows

Théorème 2.2 *Let $f \in C^3(S)$. Then*

$$\begin{aligned} f(x) - L[f](x) &= \frac{1}{2}\sum_{i=0}^{d}\left(\int_0^1 D^3_{\nu_i-x} f(x + t(\nu_i - x))(1-t)^2dt\right)\varphi_i(x) + \\ & \sum_{i=0}^{d}\sum_{r=0,r\neq i}^{d}\left(\int_0^1 D_{\nu_i-\nu_r}D^2_{x-\nu_i} f(x + t(\nu_i - x))(1-t)dt\right)\phi_{i,r}(\lambda(x)) \end{aligned}$$

It is interesting to mention that interpolation by L is preserved by H.

Proposition 2.1 *For all $f \in C^4(S)$, we have*

1. $H[f](\nu_i) = f(\nu_i), \forall i = 0, \cdots, d.$
2. $\nabla H(f](\nu_i) = \nabla f(\nu_i), \forall i = 0, \cdots, d.$
3. $H[P] \equiv P$ *, for all cubic polynomial P.*

References

[1] A. Guessab, G. Schmeisser, *harp error estimates for interpolatory approximation on convex polytopes*, SIAM J. Numer. Anal. **43** (2006) 909-923.

[2] A. Guessab, O. Nouisser, G. Schmeisser, *A Multivariate approximation by a combination of modified taylor polynomials*, J. Comput. Appl. Math. **196** (2006) 162-179.

[3] Thomas F. Sturm, *A unique multivariate hermite interpolant on the simplex*, J. Math. Anal. Appli.**191**(1995) 101-117.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Discrete orthogonal polynomial technique for parameter determination in MOSFETs transitors

D. Barrera[1], M. J. Ibáñez[1], A. M. Roldán[2], J. B. Roldán[2] and R. Yáñez[1]

[1] *Department of Applied Mathematics, University of Granada*

[2] *Department of Electronics, University of Granada*

emails: {`dbarrera`, `mibanez`, `amroldan`, `jroldan`, `ryanez`}`@ugr.es`

Abstract

In this communication, we develop a technique, based on discrete orthogonal polynomials, to determine the straight line portions in a data cloud. A practical case to determine basic parameters of a MOSFET transistor is shown.

Key words: discrete orthogonal polynomials, straight line portion, MOSFET
MSC 2000: 33C45,42C10, 65D10, 82D37

1 Introduction

Transistors, and in particular MOSFETs (Metal Oxide Semiconductor Field Effect Transistors), are the most used basic building blocks of integrated circuits (ICs) [1]. The complexity of current chips makes essential their accurate characterization to use them for circuit design purposes. For each generation of transistors the main electrical features have to be modeled in order to reproduce them as a function of the voltages differences applied between their terminals. The models (usually known as compact models) consist of a set of analytical equations and a set of parameters to include in those equations. A different set of parameters is used for each fabrication technology. These models are used in TCAD circuit simulation tools and also for hand-calculations used at the first stages of circuit design.

The extraction of the parameters of new technologies is essential since the capacities of circuit designers are dependant on the accuracy of model parameters that in many cases are linked to important physical effects.

Each parameter is obtained in a different way. However, few of them share some features in common, at least from the numerical viewpoint. In this respect, several parameters are obtained by means of extrapolation methods (for example threshold voltage calculation [1]), linear regression (determination of the body factor [1]), slope calculations (extraction of the DIBL parameter[1]), etc. In all these procedures, the determination of portions of curves that can be approximated by a straight line is crucial. In this work we just deal with this issue trying to shed light by means of advanced numerical techniques.

We have developed a method to determine the number of straight line portions contained in a curve in an automatic manner. The algorithm developed, based on discrete orthogonal polynomials, can be used for parameter extraction purposes. It consist on the isolation of straight line portions in experimental or simulated data and the determination of the slope of those curve sections to calculate one or more parameters of a compact model.

This work is organized as follows: in Section 2 we briefly describe the basic properties of the discrete orthogonal Chebyshev polynomials that we will use in this work, in Section 3 we report and discuss on the numerical procedure employed, and finally, the main conclusions are drawn in Section 4.

2 Discrete Orthogonal Chebyshev polynomials: basic properties

In these section we will briefly describe basic properties of discrete orthogonal Chebyshev polynomials that will be used in this work.

The discrete Chebyshev orthonormal polynomials are defined as the following hypergeometric series [2, 3, 4]:

$$\tilde{t}_n(x;L) = \frac{1}{n!}\sqrt{\frac{(2n+1)(L-n-1)!}{(L+n)!}}\sum_{k=0}^{n}(-1)^k\frac{n!}{k!(n-k)!}(x-k-L+1)_n(x-k+1)_n, \quad (1)$$

where $L \in \mathbb{N}$ is a parameter, $x \in \mathbb{N}$, $0 \le x \le L-1$, is the variable, and $n \in \mathbb{N}$, $0 \le n \le L-1$, is the degree of the polynomial.

They verify the three term recurrence relation

$$\begin{aligned}\tilde{t}_{n+1}(x;L) &= \frac{d_n}{d_{n+1}}(x-\beta_n)\tilde{t}_n(x;L) - \gamma_n\frac{d_{n-1}}{d_{n+1}}\tilde{t}_{n-1}(x;L)\\ &= \frac{1}{\alpha_n}(x-\beta_n)\tilde{t}_n(x;L) - \frac{\alpha_{n-1}}{\alpha_n}\tilde{t}_{n-1}(x;L) \quad (2)\end{aligned}$$

with

$$\begin{aligned}\alpha_n &= \frac{n+1}{2}\sqrt{\frac{L^2-(n+1)^2}{(2n+1)(2n+3)}},\\ \beta_n &= \frac{L-1}{2}\end{aligned}$$

and they are orthogonal with respect to the scalar product:

$$\langle \tilde{t}_n(\cdot, L), \tilde{t}_m(\cdot, L) \rangle = \sum_{x_i=0}^{L-1} \tilde{t}_n(x_i; L) \tilde{t}_m(x_i; L) = \delta_{n,m} \tag{3}$$

3 Numerical procedure

In this section we will explain the procedure followed to isolate de straight line portions in the data (experimental or simulated).

Let's considered a set of discrete data $\{x_i, y_i\}$, $i = 1, \dots N$, with x_i equispaced, i.e. $x_{i+1} - x_i = h$ (constant). Our purpose is to determine subsets of data $\{x_i, y_i\}$, $i = i_m, \dots, i_M$ forms a straight line. The procedure can be divided in two parts: i) Given a subset of data, determine if these data forms a straight line and ii) Selection of subsets. Let's analyze the two parts:

3.1 Checking of straight lines

Let's take a subset of data $\{x_i, y_i\}$, $i = i_1, \dots, i_M$ and consider the scalar products

$$r_n = \langle y, \tilde{t}_n(\cdot, M) \rangle = \sum_{i=0}^{M-1} y_i \tilde{t}_n(i, M)$$

The data can be obtained from a linear combination of the polynomials $\tilde{t}_n$:

$$y_i = \sum_{j=0}^{\infty} c_j^i \tilde{t}_j(x_i, M)$$

so the scalar products r_n takes the value

$$r_n = c_n$$

where we have used the orthonormality of the polynomials t_n. So, if the data y_i forms a straight line, we will have that all the scalar products r_n will be 0, except r_0 and/or r_1. In practice, we will consider only the scalar products r_n for $n = 0, \dots, n_{\max}$, with $n_{\max}$ given. Then, we check if all the coefficients r_n, $n = 0, \dots, n_{\max}$, vanish except r_0 and/or r_1. Only in this case, we will accept the data as a straight line and n. In practice we will consider that a coefficient r_n vanish if $|r_n| < \epsilon$. Of course, it can happens that the data y_i comes from a combination of t_0, t_1 and polynomials with higher degrees than $n_{\max}$, but, considering that we can choose $n_{\max}$ sufficiently large and that out data are experimental and, so, affected by measuring errors, this is highly improvable.

3.2 Determination of straight subsets

We start considering a subset of data with, at least, 4 points, as with 2 points we always have a straight line, but we can choose this initial number of data points arbitrarily from 3. So, we initially begin with $i_1 = 1$ and $i_M = 4$. We then check if it forms a straight line. If it does, we begin a binary search for larger subsets of straight line data between $i_M = 4$ and N (the total number of data points), ending when it detects the maximal subset of straight line data points. After that we set $i_i = i_M + 1$ (the last point of the previous straight line plus 1) and $i_M = i_1 + 4$ and start the procedure again. If the initial data points (i_1 and $i_1 + 4$ doesn't forms a straight line we increase each by 1 ($i_1 \to i_1 + 1$ and $i_M \to i_M + 1$) and start the procedure again. We stop the whole process when there are not enough data to consider a straigh line.

Once we have the subsets of straight line data, we compute the straight lines themselves, as

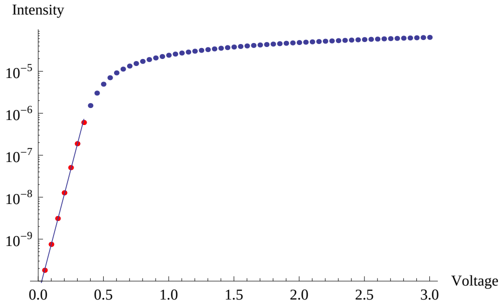
$$p(x) = c_0 t_0(x, M) + c_1 t_1(x, M)$$

This procedure can be straightforwardly generalized to detect substes of data with polynomial degree higher than 1.

4 Practical case

Here we will use the previously discused techniques to find the stright line subset of data in a MOSFET device with the following characteristics: $L = 0.3\,\mu\text{m}$, NA (channel doping concentration at the oxide surface level) $= 2.8\ 10^{16}\,\text{cm}^{-3}$, and oxide thickness $Tox = 8.7\,\text{nm}$. A N^+ polysilicon gate was used

Applying our technique to the simulated data for this MOSFET we find a straight line subset of data in the range $[0.0V, 0.35V]$



This straight line can be easily computed, finding

$$p(x) = 10^{11.9384x - 10.3234}$$

Aknowledgments

This work was partially carried out within the framework of Research Project P08-TIC-3580 supported by the Junta de Andalucía.

References

[1] N. Arora, *MOSFET modelling for VLSI simulation. Theory and practice*, reprinted by World Scientific, 2007.

[2] A.F. Nikiforov & V.B. Uvarov, Special Functions of Mathematical Physics. Birkhäuser Verlag, Basilea, 1988

[3] A.F. Nikiforov, S.K. Suslov & V.B. Uvarov, Classical Orthogonal Polynomials of a Discrete Variable. Springer Series in Computational Physics. Springer-Verlag, Berlin, 1991.

[4] R. Álvarez-Nodarse, Polinomios hipergeométricos clásicos y q-polinomios. Monografías del Seminario Matemático "García de Galdeano".Vol. 26. Prensas Universitarias de Zaragoza, Zaragoza, Spain, 2003.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Conservation Law Construction via Mathematical Fluctuation Theory for Exponentially Anharmonic, Symmetric, Quantum Oscillator

Semra Bayat[1] **and Metin Demiralp**[1]

[1] *Informatics Institute, Computational Science and Engineering Program, İstanbul Technical University*

emails: `bayat@itu.edu.tr`, `metin.demiralp@gmail.com`

Abstract

This work focuses on the point that how conservation laws amongst the expectation values can be constructed. We use the expectation value dynamics of quantum systems and the mathematical fluctuation concept. The fluctuation free equations match the classical mechanical equations while the introduction of the fluctuations increases the number of the ODEs even though the added ODEs are linear in fluctuations. We have shown that space extension creates space contraction laws while each different order set of fluctuation creates interfluctuation conservation laws. The familiar laws like the energy conservation still exist but there is a need to have fluctuation related terms.

Key words: Expectation Values, Quantum Expected Value Dynamics, Fluctuations, Conservation Laws.

1 Introduction

We concern with the conservation law construction for quantum systems by using mathematical fluctuation theory in this work. Our focus is taken as the quantum system whose potential contains a symmetric exponential function which shows the harmonic oscillator behaviour at the zero limit of the scaling parameter in its argument. The Schrödinger equation [1] for this system can be explicitly given through the following equalities

$$i\hbar\frac{\partial\psi}{\partial t}=\widehat{H}\psi,\qquad \psi\left(\mathbf{x},0\right)=\psi_0\left(x\right) \tag{1}$$

$$\widehat{H} = \frac{1}{2\mu}\widehat{p}^{\,2} + \alpha\left(e^{\frac{k}{2}\widehat{q}^{\,2}} - \widehat{I}\right) \tag{2}$$

$$\widehat{p}f(x) \equiv -i\hbar\frac{\partial f(x)}{\partial x}, \qquad \widehat{q}f(x) \equiv xf(x) \tag{3}$$

where $\widehat{I}$, $\widehat{p}$, $\widehat{q}$, and, $\widehat{H}$ denote the identity, momentum, position, and, Hamilton operators of the system under consideration. The system has one degree of freedom and describes the motion of a particle which can move on a straight line by definition. μ, α, and, k stand for the mass, potential amplitude, and, anharmonicity constant which are parameters to be used for system identification. $\hbar$ denotes the universal reduced Planck constant while the function $\psi_0\left(x\right)$ symbolizes the initial form of the wave function. All these enable us to rewrite the system's Schrödinger equation in a single formula explicitly as follows

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = -\frac{\hbar^2}{2\mu}\frac{\partial^2\psi(x,t)}{\partial x^2} + \alpha\left(e^{\frac{k}{2}x^2} - 1\right)\psi(x,t), \qquad \psi(x,0) = \psi_0\left(x_0\right) \tag{4}$$

which is a parabolic partial differential equation describing the system's evolution. The general tendency is to find the wave function by solving this equation and then to evaluate the expectation values of certain operators corresponding to some observable entities. Hence the wave function is in fact an intermediate tool and necessitates the solution of a partial differential equation (PDE) [2]. This is not an easy task even though the present PDE seems to be rather easy to be solved, since the increasing degree of freedom in the system at the focus complicates the problem of solving PDE because of the increasing dimensionality. This important fact urged us to skip the evaluation of the wave function and construct the ODEs over the expectation values we need to evaluate [3–5]. We call this approach "Quantum Expectation Value Dynamics" because of its nature. We will use this approach for the evaluation of certain expectation values for our target system here. However, our main purpose is beyond this. We want to construct certain conservation laws like the energy conservation of the mechanics. We seek the relations amongst the expectation values of certain operators such that they do not change functional structure during the evolution of the system. We take the first steps in the next section and then we focus on mathematical fluctuation concept on the continuation of the study [6–9]. Relevant fluctuation equations are constructed in the following sections.

2 Expectation Value Dynamics on an Extended Space

The exponential function structure in the potential of the target system above complicates the formulation since it does not have a finite term representation in powers of its argument. On the other hand, despite its highly complicated singularity at infinity the exponential function has a very specific and pleasant property. Its derivative with respect to its argument is same as itself. Exponential function does not linearly depend on certain finite number of powers of its argument even though there is a functional dependence amongst them. This brings the idea of considering the exponential function as if an additional independent variable and therefore to work in an extended space [10, 11]. Hence, we can start with the definition

$$\widehat{r}f(x) \equiv \left(e^{\frac{k}{2}x^2} - 1 \right) f(x) \tag{5}$$

where $\widehat{r}$ is apparently an algebraic multiplication operator. This enables us to write the following Poisson brackets

$$\left\{ \widehat{H}, \widehat{p} \right\} \equiv \frac{i}{\hbar} \left[\widehat{H}, \widehat{p} \right] = -\alpha k \widehat{q} \left(\widehat{r} + \widehat{I} \right) \tag{6}$$

$$\left\{ \widehat{H}, \widehat{q} \right\} \equiv \frac{i}{\hbar} \left[\widehat{H}, \widehat{q} \right] = \frac{1}{\mu} \widehat{p} \tag{7}$$

$$\left\{ \widehat{H}, \widehat{r} \right\} = \frac{k}{2\mu} \widehat{pq} \left(\widehat{r} + \widehat{I} \right) + \frac{k}{2\mu} \left(\widehat{r} + \widehat{I} \right) \widehat{qp} \tag{8}$$

These Poisson brackets are in fact the core agents of the temporal derivatives of certain operator expectation values as the following analysis shows.

The explicit expectation value definition of a given operator $\widehat{o}$ is as follows

$$\left\langle \widehat{o} \right\rangle (t) \equiv \int_V dV \psi \left(x, t \right)^* \widehat{o} \psi \left(x, t \right) \tag{9}$$

the simple temporal differentiation of this equality produces

$$\frac{d \left\langle \widehat{o} \right\rangle (t)}{dt} = \left\langle \left\{ \widehat{H}, \widehat{o} \right\} \right\rangle (t) \tag{10}$$

where

$$\left\{ \widehat{H}, \widehat{o} \right\} \equiv \frac{i}{\hbar} \left(\widehat{H}\widehat{o} - \widehat{o}\widehat{H} \right). \tag{11}$$

These allow us to obtain

$$\frac{d \left\langle \widehat{p} \right\rangle}{dt} = -\alpha k \left\langle \widehat{q} \left(\widehat{r} + \widehat{I} \right) \right\rangle, \quad \left\langle \widehat{p} \right\rangle (0) = \left\langle \widehat{p} \right\rangle_0 \tag{12}$$

$$\frac{d \left\langle \widehat{q} \right\rangle}{dt} = \frac{1}{\mu} \left\langle \widehat{p} \right\rangle, \quad \left\langle \widehat{q} \right\rangle (0) = \left\langle \widehat{q} \right\rangle_0 \tag{13}$$

$$\frac{d \left\langle \widehat{r} \right\rangle}{dt} = \frac{k}{2\mu} \left\langle \widehat{pq} \left(\widehat{r} + \widehat{I} \right) \right\rangle + \frac{k}{2\mu} \left\langle \left(\widehat{r} + \widehat{I} \right) \widehat{qp} \right\rangle, \quad \left\langle \widehat{r} \right\rangle (0) = \left\langle \widehat{r} \right\rangle_0 \tag{14}$$

The right hand sides of the latest three equations except the second one are expectation values of certain operators which are linearly independent from the left hand side operators. Hence, we can not consider these three equations as a complete set of ODEs. We need to construct further equations. We do this in the next section by using fluctuation concept.

3 Conservation Laws at Fluctuationlessness Limit

Let us define the following fluctuation operators somehow describing the deviations from the expectation values for $\widehat{p}$, $\widehat{q}$ and $\widehat{r}$

$$\widehat{\phi}_p = \widehat{p} - \left\langle \widehat{p} \right\rangle \widehat{I}, \qquad \widehat{\phi}_q = \widehat{q} - \left\langle \widehat{q} \right\rangle \widehat{I}, \qquad \widehat{\phi}_r = \widehat{r} - \left\langle \widehat{r} \right\rangle \widehat{I}, \tag{15}$$

then

$$\widehat{p} = \left\langle \widehat{p} \right\rangle \widehat{I} + \widehat{\phi}_p, \qquad \widehat{q} = \left\langle \widehat{q} \right\rangle \widehat{I} + \widehat{\phi}_q, \qquad \widehat{r} = \left\langle \widehat{r} \right\rangle \widehat{I} + \widehat{\phi}_r. \tag{16}$$

The potential function defined as

$$V\left(\widehat{q}\right) = \alpha \left(e^{\frac{k}{2}\widehat{q}^2} - \widehat{I} \right) \tag{17}$$

can be rewritten by using fluctuation operators defined above

$$\begin{aligned} V\left(\widehat{q}\right) = V\left(\left\langle \widehat{q} \right\rangle \widehat{I} + \widehat{\phi}_q \right) &= \alpha \left(e^{\frac{k}{2}\left(\left\langle \widehat{q} \right\rangle \widehat{I} + \widehat{\phi}_q \right)^2} - \widehat{I} \right) \\ &= \sum_{j=0}^{\infty} \frac{\left\{ \alpha \left(e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} - \widehat{I} \right) \right\}^{(j)}}{j!} \widehat{\phi}_q^j \\ &= \alpha \left(e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} - \widehat{I} \right) \widehat{I} + \alpha k \left\langle \widehat{q} \right\rangle e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} \widehat{\phi}_q + \sum_{j=2}^{\infty} \frac{\left\{ \alpha \left(e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} - \widehat{I} \right) \right\}^{(j)}}{j!} \widehat{\phi}_q^j \end{aligned} \tag{18}$$

The expected value of the potential function $V\left(\widehat{q}\right)$ can be evaluated as

$$\begin{aligned} \left\langle V\left(\widehat{q}\right) \right\rangle &= \alpha e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} + \sum_{j=2}^{\infty} \frac{\left(\alpha e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} \right)^{(j)}}{j!} \left\langle \widehat{\phi}_q^j \right\rangle \\ &= \alpha e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} + \sum_{j=1}^{\infty} \frac{\left(\alpha e^{\frac{k}{2}\left\langle \widehat{q} \right\rangle^2} \right)^{(j+1)}}{(j+1)!} \left\langle \widehat{\phi}_q^{j+1} \right\rangle \end{aligned} \tag{19}$$

where we have taken into consideration that the expected value of any one of fluctuation operators defined above vanishes. Despite this vanishing feature the expectation values of the higher degree terms formed by the products of powers of the fluctuation operators do not vanish. We call those entities "Fluctuations". The j^{th} order fluctuations which can be derived from momentum and position operator expectation values can be defined as follows

$$\varphi_{j,k} = \left\langle \frac{1}{2} \left(\widehat{\phi}_p^k \widehat{\phi}_q^{j+1-k} + \widehat{\phi}_q^{j+1-k} \widehat{\phi}_p^k \right) \right\rangle \qquad j = 1, 2, ... \quad k = 0, 1, ..., j+1. \tag{20}$$

If we now neglect all fluctuation terms in the potential function expansion and in the right hand side expansion of the expectation value ODE for the operator $\widehat{r}$ then we can write

$$\begin{aligned}
\frac{d\left\langle\widehat{p}\right\rangle}{dt} &= -\alpha k\left\langle\widehat{q}\right\rangle\left(\left\langle\widehat{r}\right\rangle+\widehat{I}\right) && (21)\\
\frac{d\left\langle\widehat{q}\right\rangle}{dt} &= \frac{1}{\mu}\left\langle\widehat{p}\right\rangle && (22)\\
\frac{d\left\langle\widehat{r}\right\rangle}{dt} &= \frac{k}{\mu}\left\langle\widehat{p}\right\rangle\left\langle\widehat{q}\right\rangle\left(\left\langle\widehat{r}\right\rangle+\widehat{I}\right). && (23)
\end{aligned}$$

It is not hard to get the following relation from the second and third equations of this set

$$\frac{k}{2}\left\langle\widehat{q}\right\rangle^2-\ln\left(\widehat{I}+\left\langle\widehat{r}\right\rangle\right)=c_1 \qquad (24)$$

where c_1 is an arbitrary constant for this moment. This is the reflection of the space extension realized by defining $\widehat{r}$. That equation is in fact an algebraic identification and we can enforce its validity for the initial expectation value quite naturally. This makes c_1 vanishing and we can write the following equality at the end

$$\left\langle\widehat{r}\right\rangle=\mathrm{e}^{\frac{k}{2}\left\langle\widehat{q}\right\rangle^2}-\widehat{I} \qquad (25)$$

This is somehow the reflection of the algebraic relation between $\widehat{r}$ and the position operator to the relation between these two operators' expectation values. In other words, this relation takes us back from the extended space to the original space. Hence we can call this relation "Space Contraction Law".

The utilization of (25) in the first one of triple equation set in (23) and combining the resulting equation with the second equation of that set takes us to a couple of ODEs in momentum and position operator. Due to the structure of this couple of ODEs we can integrate them to the following relation

$$\frac{1}{2\mu}\left\langle\widehat{p}\right\rangle^2+\alpha\left(\mathrm{e}^{\frac{k}{2}\left\langle\widehat{q}\right\rangle^2}-\widehat{I}\right)=c_2 \qquad (26)$$

where c_2 is an arbitrary constant for the moment. It in fact corresponds to the total energy of the system under consideration and this relation states that the expectation value of the system Hamiltonian is conserved for all time instances. To evaluate the conserved total energy value it is sufficient the evaluate the expectation values at the initial moment where the wave function is given. So the second law given by (26) is the "Energy Conservation Law" for the system under consideration.

4 Conservation Laws in First Fluctuation Point of View

In the previous sections we have focused on the expectation value dynamics in an extended space to show the role of space extension explicitly. In this section we focus on the case where no space extension is realized. Hence, the operator $\widehat{r}$ will be undefined and all the system properties will be described certain entities depending on the momentum and position operators only. This time we are going to omit again all fluctuations except the ones with first order. The approximate form of the system Hamiltonian's expectation value can be given as follows for these cases

$$\langle H\rangle = \frac{1}{2\mu}\left\langle \widehat{p}\right\rangle^2 + \alpha \mathrm{e}^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2} + \frac{1}{2\mu}\varphi_{1,2} + \frac{\alpha k}{2}\left(\widehat{I} + k\left\langle \widehat{q}\right\rangle^2\right)\mathrm{e}^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2}\varphi_{1,0} \tag{27}$$

This is in fact the first order fluctuation involving approximate form of the overall energy conservation. The expectation value dynamics of the momentum and position operators within the first order fluctuation approximation can be written as follows

$$\begin{aligned} \frac{d\left\langle \widehat{p}\right\rangle}{dt} &= -\alpha k\left\langle \widehat{q}\right\rangle \mathrm{e}^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2} - \frac{\alpha k}{2}\left(\widehat{I} + k\left\langle \widehat{q}\right\rangle^2\right)\mathrm{e}^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2}\varphi_{1,0} && (28)\\ \frac{d\left\langle \widehat{q}\right\rangle}{dt} &= \frac{1}{2\mu}\left\langle \widehat{p}\right\rangle. && (29)\end{aligned}$$

These equations are somehow incomplete since they contain unknown fluctuations which are temporal functions. Hence we need to add certain ODEs to get a complete set of ODEs. To this end we can start with the following definitions

$$\varphi_{1,0} = \left\langle \widehat{\phi}_q^2\right\rangle, \qquad \varphi_{1,1} = \left\langle \frac{1}{2}\left(\widehat{\phi}_p\widehat{\phi}_q + \widehat{\phi}_q\widehat{\phi}_p\right)\right\rangle, \qquad \varphi_{1,2} = \left\langle \widehat{\phi}_p^2\right\rangle \tag{30}$$

These definitions can be used to construct ODEs for these entities. For this action we can employ the Poisson brackets of certain operators and some operator algebra together with standard algebra. By skipping the intermediate derivation stages we can obtain

$$\begin{aligned} \frac{d\varphi_{1,0}}{dt} &= \frac{2}{\mu}\varphi_{1,1} && (31)\\ \frac{d\varphi_{1,1}}{dt} &= -\left(\alpha k e^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2} + \alpha k^2\left\langle \widehat{q}\right\rangle^2 \mathrm{e}^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2}\right)\varphi_{1,0} + \frac{1}{\mu}\varphi_{1,2} && (32)\\ \frac{d\varphi_{1,2}}{dt} &= -2\left(\alpha k e^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2} + \alpha k^2\left\langle \widehat{q}\right\rangle^2 e^{\frac{k}{2}\left\langle \widehat{q}\right\rangle^2}\right)\varphi_{1,1} && (33)\end{aligned}$$

It is possible to find a conserved entity whose value does not change during the system's evolution. It can be constructed from the equations of the latest triple set of ODEs above. We can conclude

$$\varphi_{1,1}^2 - \varphi_{1,0}\varphi_{1,2} = c_3 \tag{34}$$

where c_3 stands for a constant which is arbitrary for this moment. It can be evaluated by using the initial expectation values via the initial wave form. This relation is conserved for all time instances of the evolution and we call it “First Order Fluctuation Conservation Law”. This is somehow the fluctuation reflection of the energy conservation law which relates the expectation values of the momentum and position operators.

The other conservation law is again energy conservation law and is based on the invariance of the Hamiltonian’s expectation value which depends on the first order fluctuations as we have mentioned at the beginning of this section. This completes the investigation for the first order fluctuation preserving case.

5 Concluding Remarks

The purpose of this work has been the construction of conservation laws in the framework of the mathematical fluctuation theory. We have used the expectation value dynamics to this end. We also have shown the role of the space extension in facilitating the analysis. Because of the highly probabilistic nature of the quantum dynamical problems we needed to use the fluctuation concept as the expectation values of the powers of certain deviation (fluctuation) operators. We have reobtained the well known “Energy Conservation Law” for no fluctuation and first order fluctuations. The cases where space extension has been used certain “Space Contraction Laws” have been obtained. We have interpreted them as going back to the original space from the extended space. The additions fluctuation concept has arisen the conservation law(s) amongst the fluctuations. We have limited our investigations to the cases where at most first order fluctuations are taken into consideration. All these analyses can be extended to much higher order fluctuation containing cases without any remarkable conceptual difficulty even though the volume of the manipulations and intermediate steps increases pretty much.

References

[1] P. A. M. Dirac, *The Fundamental Equations of Quantum Mechanics*, Proc. R. Soc. Lond. A **109** (1925) 642–653.

[2] H. J. W. Muller-Kirsten, *Introduction to Quantum Mechanics: Schrodinger Equation And Path Integral*, World Scientific Publishing Company, Singapore, 2006.

[3] M. Demiralp, *Determination of the Quantum Motion of the One Dimensional Harmonic Oscillator via Expectation Value Evolutions*, Bull. Tech. Univ. Istanbul **47**(4) (1994) 357.

[4] M. Demiralp, *Quantum Mechanical Matrix Ordinary Differential Equations and Their Solutions by Characteristic Evolutions*, Proceeding of MMACTEE'09, Vouliagmeni, Athens (2009) 657–662.

[5] E. Meral ve M. Demiralp, *Determination Of The External Field Amplitude And Deviation Parameter Through Expectation Value Based Quantum Optimal Control Of Multiharmonic Oscillators Under Linear Control Agents*, J. Math. Chem. **46** (2009) 834–852.

[6] M. Demiralp, *Determination of Quantum Expectation Values Via Fluctuation Expansion*, Lecture Series on Computer and Computational Sciences, Selected Papers from Int. Conf. of Comput. Methods in Sci. and Eng. (ICCMSE 2005), Loutraki, Greece, **4A** (2005) 146–149.

[7] M. Demiralp, *A Fluctuation Expansion Method for the Evaluation of a Function's Expectation Value*, Proc. of the Int. Conf. on Numerical Analysis and Applied Mathematics (ICNAAM 2005), Rhodes, Greece (2005) 711–714

[8] N. Altay ve M. Demiralp, *Numerical Solution of Ordinary Differential Equations by Fluctuationlessness Theorem*, J. Math. Chem. **47**(4) (2010) 1323–1344.

[9] M. Ayvaz ve M. Demiralp, *A Fluctuation Analysis at the Classical Limit for the Expectation Dynamics of a Single Quartic Quantum Anharmonic Oscillator*, AIP Conf. Proc. **1281** (2010) 1950–1953.

[10] S. Üsküplü ve M. Demiralp, *Univariate Integration via Space Extension Based No Fluctuation Approximation*, AIP Conf. Proc. **1048** 566–569.

[11] S. Üsküplü ve M. Demiralp, *Conversion of PDEs to Certain Universal and Easily Handleable Forms Via Space Extension*, Proceedings of the 1st WSEAS Int. Conf. on Multivariate Analysis and its Application in Science and Engineering (MAASE'08) (2008) 179–182.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Optimal control of a linear and unbranched chemical process with n steps: the quasi-analytical solution

L. Bayón[1], J.M. Grau[1], M.M. Ruiz[1] and P.M. Suárez[1]

[1] *Department of Mathematics, University of Oviedo, Spain*

emails: bayon@uniovi.es, grau@uniovi.es, mruiz@uniovi.es, pedrosr@uniovi.es

Abstract

In this paper we present a method to solve a constrained optimal control problem to calculate the optimal enzyme concentrations in a chemical process by considering the minimization of the transition time. The method, based on Pontryagin's Maximum Principle, allows us to obtain, in an almost exclusively analytical way, the generalized solution of an n-step system with an unbranched scheme and bilinear kinetic models.

Key words: Optimal Control, Chemical Process, Pontryagin's Principle
MSC 2000: 49J30, 49M05, 92E20, 80A30, 92C40

1 Introduction

In this paper we present an optimal control problem that arises when metabolic chemical processes are considered. Within this context, one of the most important problems is the study of enzyme concentrations. Our work focuses on dynamic optimization, studying the problem of minimizing the transition time during which the substrate is converted into the product.

Let us consider the following (unbranched) reaction chain of n irreversible reactions 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 is the substrate concentration (starting reagent), p the concentration of the final product, x_i $(i = 2, \ldots, n)$ the concentration of the intermediate compounds, and u_i $(i = 1, \ldots, n)$ the concentration of the enzyme catalyzing the i-th reaction.

For the dynamic case, the aim is to solve the problem analytically and numerically. An explicit solution for the simplest case, i.e. $n = 2$, can be found in [1]. For longer pathways, i.e. $n > 2$, the aforementioned authors solved the optimization problem numerically. An interesting study is presented in [2] in which the solution is obtained quasi-analytically, though with the constraint of considering only the case of $n = 3$ with two intermediate compounds. An interesting theoretical result is presented in [3] for the general case of n steps: the optimal enzyme concentration profile is of the "bang-bang" type (a well-known concept in the framework of optimal control which implies that the solution switches between 0 and the maximal level), except in the last interval. Other qualitative considerations of the solution are also presented, but not the analytical solution.

In this paper we shall substantially extend the theoretical analysis of [2] and [3], presenting the quasi-analytical solution for the more general case of n steps. The paper is organized as follows. Section 2 presents the statement of the problem. In Section 3 we carry out a calculation based on Pontryagin's Maximum Principle. Finally, we present the conclusions drawn.

2 Statement of the Problem

The optimization of enzyme concentrations in metabolic pathways can be calculated using the optimality criterion of minimizing the time period during which an essential product is generated. [1] and [2] assumed bilinear (linear in the metabolite concentrations, x_i, and linear in the enzyme concentrations, u_i) and irreversible rate laws. [3] used a more general model: the rate laws are only linear in the u_i, and some assumptions are made about the behaviour of x_i. In this paper we use a bilinear kinetic model to solve the problem analytically, likewise assuming that the enzymes can be switched on and off instantaneously. For simplicity's sake, we employ normalized quantities. Enzyme levels are divided by the maximum total enzyme concentration, and substrate, intermediate and product levels by the initial substrate concentration.

Our goal is to convert substrate x_1 into product p as fast as possible. Several cost functions may be considered. In [3], combined optimization of the time taken to reach the new steady state and a measure of enzyme usage is considered. In this paper, we use the *transition time,* τ, as defined in [4], which is likewise used in [1] and [2]. Thus, the objective function of the optimization problem may be defined as:

$$\min_{u_1,\dots u_n} \tau = \min_{u_1,\dots u_n} \int_0^\infty \frac{1}{x_1(0)}(x_1(0) - p(t))dt$$

Due to normalization, $x_1(0) = 1$, and the conservation relation:

$$x_1(t) + x_2(t) + \ldots + x_n(t) + p(t) = 1, \ \forall t \geq 0$$

the objective function can be written as:

$$\min_{u_1,\ldots u_n} \tau = \min_{u_1,\ldots u_n} \int_0^\infty (x_1(t) + x_2(t) + \ldots + x_n(t))dt \tag{2}$$

where the concentrations $x_1, x_2, \ldots, x_n$ are the state variables (p is eliminated) and the concentration of enzymes $u_1, u_2, \ldots, u_n$ are the control variables.

The model of the reactions in (1) can then be described by the set of differential equations (see [1] and [2]):

$$\begin{cases} \dot{x}_1 = -k_1 u_1 x_1 & x_1(0) = 1 \quad x_1(t) \geq 0 \\ \dot{x}_2 = k_1 u_1 x_1 - k_2 u_2 x_2 & x_2(0) = 0 \quad x_2(t) \geq 0 \\ \dot{x}_3 = k_2 u_2 x_2 - k_3 u_3 x_3 & x_3(0) = 0 \quad x_3(t) \geq 0 \\ \cdots & \\ \dot{x}_n = k_{n-1} u_{n-1} x_{n-1} - k_n u_n x_n & x_n(0) = 0 \quad x_n(t) \geq 0 \end{cases} \tag{3}$$

where, for the sake of simplicity, we shall assume equal catalytic efficiencies of the enzymes ($k_i = k = 1$). As an initial condition, for $t = 0$, we shall consider the concentrations of the intermediate compounds and of the product to be equal to zero. Finally, we shall consider the concentrations of the compounds, x_i, as well as those of the enzymes, u_i, to be positive limited quantities and, after normalization, that the upper bound on the enzymatic concentration is 1. Hence, $(u_1(t), ..., u_n(t)) \in \Omega$, being:

$$\Omega = \{\mathbf{u} = (u_1(t), ..., u_n(t)) \in \mathbb{R}^n \mid u_1 \geq 0, \ldots u_n \geq 0;\ u_1 + \ldots + u_n \leq 1\} \tag{4}$$

We have thus stated an optimal control problem.

3 Optimization

In this section we present the solution to the optimal control problem defined in the previous section using Pontryagin's Minimum Principle (PMP) [5]. In our case, as regards the control appearing linearly in the Hamiltonian function H:

$$H = x_1 + x_2 + \cdots + x_n + \lambda_1(-u_1 x_1) + \lambda_2(u_1 x_1 - u_2 x_2) + \cdots + \lambda_n(u_{n-1} x_{n-1} - u_n x_n)$$

when H is minimized w.r.t. the control variables:

$$\min_{\mathbf{u}} H = \min_{\mathbf{u} \in \Omega} \{-\mu_1 u_1 - \mu_2 u_2 - \cdots - \mu_n u_n\}; \quad \begin{cases} \mu_1 = (\lambda_1 - \lambda_2) x_1 \\ \mu_2 = (\lambda_2 - \lambda_3) x_2 \\ \vdots \\ \mu_{n-1} = (\lambda_{n-1} - \lambda_n) x_{n-1} \\ \mu_n = \lambda_n x_n \end{cases} \tag{5}$$

it is shown that control u_i will be activated when the *switching function* μ_i reaches its maximum value. If u_i switches between its upper and lower bounds only at isolated points in time, then the optimal control is said to be a bang-bang type control. The times are called *switching times*. We shall obtain the optimal solution constructively by intervals, starting from $t = 0$ and concatenating the results. The fundamental result to obtain may be summarized as follows:

Proposition 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:*

$$[0, t_1) \cup [t_1, t_2) \cup \cdots \cup [t_{n-2}, t_{n-1}) \cup [t_{n-1}, \infty)$$

such that the optimal profile of the i-th enzyme satisfies:

$$u_i^*(t) = \begin{cases} 1 & \text{for} \quad t \in [t_{i-1}, t_i) \\ 0 & \text{fot} \quad t \notin [t_{i-1}, t_i) \end{cases} \; ; \; i = 1, \ldots, n-1$$

with $t_0 = 0$. *In the last interval* $(t \geq t_{n-1})$, *the solution is not of the bang-bang type.*

Interval	Concentrations	τ_i
$[0, t_1]$	$x_1(t) = e^{-t}$ $x_2(t) = 1 - e^{-t}$ $x_3(t) = 0; \ldots ; x_n(t) = 0$	t_1
$[t_1, t_2]$	$x_1(t) = e^{-t_1}$ $x_2(t) = \left(1 - e^{-t_1}\right) e^{-(t-t_1)}$ $x_3(t) = \left(1 - e^{-t_1}\right) \left(1 - e^{-(t-t_1)}\right)$ $x_4(t) = 0; \ldots ; x_n(t) = 0$	$t_2 - t_1$
$[t_2, t_3]$	$x_1(t) = e^{-t_1}$ $x_2(t) = \left(1 - e^{-t_1}\right) e^{-(t_2-t_1)}$ $x_3(t) = \left(1 - e^{-t_1}\right) \left(1 - e^{-(t_2-t_1)}\right) e^{-(t-t_2)}$ $x_4(t) = \left(1 - e^{-t_1}\right) \left(1 - e^{-(t_2-t_1)}\right) \left(1 - e^{-(t-t_2)}\right)$ $x_5(t) = 0; \ldots ; x_n(t) = 0$	$t_3 - t_2$
...	...	...
$[t_{n-2}, t_{n-1}]$	$x_1(t) = e^{-t_1}$ $x_2(t) = \left(1 - e^{-t_1}\right) e^{-(t_2-t_1)}$ $x_3(t) = \left(1 - e^{-t_1}\right) \left(1 - e^{-(t_2-t_1)}\right) e^{-(t_3-t_2)}$ $\vdots$ $x_{n-2}(t) = \left(1 - e^{-t_1}\right) \cdots (1 - e^{-(t_{n-3}-t_{n-4})}) e^{-(t_{n-2}-t_{n-3})}$ $x_{n-1}(t) = \left(1 - e^{-t_1}\right) \cdots (1 - e^{-(t_{n-2}-t_{n-3})}) e^{-(t-t_{n-2})}$ $x_n(t) = \left(1 - e^{-t_1}\right) \cdots (1 - e^{-(t_{n-2}-t_{n-3})})(1 - e^{-(t-t_{n-2})})$	$t_{n-1} - t_{n-2}$

(6)

The optimal solution is obtained analytically for the intervals $[0,t_1)\cup[t_1,t_2)\cup\cdots\cup[t_{n-2},t_{n-1})$. The value of u_i is given by Proposition 1, while the values of the concentrations $x_1,x_2,\ldots,x_n$ are given by (6). The transition times τ_i are defined by:

$$\tau_i=\int_{t_{i-1}}^{t_i}(x_1(t)+x_2(t)+x_3(t)+\cdots+x_n(t))\,dt;\ i=1,\ldots,n-1$$

In the last interval, $[t_{n-1},\infty)$, it is observed that u_n cannot be activated. Therefore, in order to calculate the solution in this last interval, we need to determine the minimum total transition time, τ. It can be seen that $\tau(t_1,t_2,\ldots,t_{n-1},u_1,u_2,\ldots,u_n)$ is given by:

$$\tau=t_{n-1}+x_1(t_{n-1})\left(\frac{1}{u_1}+\cdots+\frac{1}{u_n}\right)+x_2(t_{n-1})\left(\frac{1}{u_2}+\cdots+\frac{1}{u_n}\right)+\cdots+x_n(t_{n-1})\left(\frac{1}{u_n}\right)$$

where $x_i(t_{n-1})$ are known from (6).

To minimize τ with the condition:

$$u_1+u_2+\ldots+u_n=1$$

we apply the method of Lagrange multipliers to the augmented functional:

$$L(t_1,t_2,\ldots,t_{n-1},u_1,u_2,\ldots,u_n,\beta)=\tau+\beta(u_1+u_2+\ldots+u_n-1)$$

In order to do so, we have to solve the non-lineal system:

$$\frac{\partial L}{\partial t_1}=0;\frac{\partial L}{\partial t_2}=0;\ldots;\frac{\partial L}{\partial t_{n-1}}=0;\frac{\partial L}{\partial u_1}=0;\frac{\partial L}{\partial u_2}=0;\ldots;\frac{\partial L}{\partial u_n}=0;\frac{\partial L}{\partial \beta}=0$$

which may be done by means of any commonly used program.

It is therefore in this last step when we truly determine the switching times: $t_1,t_2,\ldots,t_{n-1}$, and the values of $u_1,u_2,\ldots,u_n$ in the last interval, $[t_{n-1},\infty)$ (in the other intervals u_i is given by Proposition 1). The problem is completely solved by calculating $x_1(t),x_2(t),\cdots,x_n(t)$ in $[t_{n-1},\infty)$ by means of the following equations:

$$\begin{cases}
x_1(t)=x_1(t_{n-1})e^{-u_1(t-t_{n-1})}\\
x_2(t)=x_2(t_{n-1})e^{-u_2(t-t_{n-1})}\\
+\frac{u_1}{u_2-u_1}x_1(t_{n-1})\left(e^{-u_1(t-t_{n-1})}-e^{-u_2(t-t_{n-1})}\right)\\
x_3(t)=x_3(t_{n-1})e^{-u_3(t-t_{n-1})}+\frac{u_2}{u_3-u_2}x_2(t_{n-1})e^{-u_2(t-t_{n-1})}\\
+\frac{u_2u_1}{u_2-u_1}x_1(t_{n-1})\left(\frac{e^{-u_1(t-t_{n-1})}}{u_3-u_1}-\frac{e^{-u_2(t-t_{n-1})}}{u_3-u_2}\right)\\
-\frac{u_2}{u_3-u_2}x_2(t_{n-1})e^{-u_3(t-t_{n-1})}\\
-\frac{u_2u_1}{u_2-u_1}x_1(t_{n-1})(\frac{1}{u_3-u_1}-\frac{1}{u_3-u_2})e^{-u_3(t-t_{n-1})}\\
\vdots
\end{cases}$$

We have thus solved the problem quasi-analytically; this last step, the calculation of the switching times, being the only one that is not carried out analytically or exactly.

4 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 transition time.

References

[1] E. Klipp, R. Heinrich, H.G. Holzhütter, *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. Oyarzún, B. Ingalls, R. Middleton, D. Kalamatianos, *Sequential activation of metabolic pathways: a dynamic optimization approach*, Bull. Math. Biol. **71(8)** (2009) 1851–1872.

[4] 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.

[5] R. Vinter, *Optimal Control, Systems & Control: Foundations & Applications.* Birkhäuser Boston, Inc., Boston, MA, 2000.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Linear and Cyclic Codes over a Finite Non Chain Ring

Aysegul Bayram[1] **and Irfan Siap**[1]

[1] *Department of Mathematics, Faculty of Arts and Sciences, Yildiz Technical University*

emails: abayram@yildiz.edu.tr, isiap@yildiz.edu.tr

Abstract

In this paper, we study linear and cyclic codes over the non chain ring $Z_p[v]/\langle v^p - v\rangle$ where p is a prime number. A distance preserving Gray map which induces a relation between codes over this ring and Z_p codes is introduced. Further, the algebraic structure of cyclic and dual codes over $Z_p[v]/\langle v^p - v\rangle$ is also studied.

Key words: Non chain ring, linear codes, cyclic codes,
MSC 2000: AMS codes (11T71,94B15,94B05)

1 Introduction

Codes over different special rings have been recently considered. These codes are interesting if they can be related to codes over finite fields via some special maps generally known as Gray maps. [8] is the first significant paper that relates codes over the residue ring of integers of size 4, i.e. $\mathbb{Z}_4$. In [8], some linear codes over $\mathbb{Z}_4$ have been mapped to binary codes which are non linear codes in general via a Gray map and some well known optimal binary codes, such as Kerdock, Preparata codes, are obtained as images under this map. Then codes over the ring F_2+uF_2, $F_2+uF_2+u^2F_2$, $F_2[u]/\langle u^s\rangle$ and their algebraic structures are studied in [2, 3, 4, 6, 10, 12]. Optimal self dual codes over the ring $F_2[v]/\langle v^2 - v\rangle$ has been studied in [5] which motivated us. Lately, some other finite rings which do not posses chain property are also considered. These rings are more difficult to deal with but they enjoy Gray maps that are interesting. Some examples on this direction are codes over the ring $Z_2[u, v]/\langle u, v\rangle$ [13, 14]. Some further study over more generalized rings of this type that include the ring $Z_2[u, v]/\langle u, v\rangle$ are being carried out pretty recently [7, 13, 14]. Cyclic codes over the ring $Z_2[v]/\langle v^2 - v\rangle$ and $Z_3[v]/\langle v^3 - v\rangle$ are studied [1, 5, 15].

In this paper, we study codes over the ring $R = Z_p[v]/\langle v^p - v\rangle$, where p is a prime number. This work is a generalization of that in [1]. In the first section we introduce the ring, its algebraic structure, ideals, units, etc. Next linear codes over R are considered. Then by defining an inner product the dual of a linear code is defined and relation to linear code and its dual is also presented. Finally, the algebraic structure of cyclic codes and their duals are presented.

2 Preliminaries

In this section we study the ring $Z_p[v]/\langle v^p - v\rangle = \{a_0 + a_1 v + ... + a_{p-1}v^{p-1} | a_0, a_1, ..., a_{p-1} \in Z_p$ and $v^p = v\}$, its units, ideal structure and its properties. We introduce a Gray map which is deduced from the Chinese Remainder Theorem and this map relates the ring R with the ring $Z_p \oplus Z_p \oplus ... \oplus Z_p$ where the number of Z_p is p.

Definition 2.1 *Let $\alpha = a_0 + a_1 v + ... + a_{p-1}v^{p-1} \in R$ and $\alpha(i) = a_0 + a_1 i + ... + a_{p-1}i^{p-1}($ mod $p)$ then $\phi(\alpha) = \phi(a_0 + a_1 v + ... + a_{p-1}v^{p-1}) = (\alpha(0), \alpha(1), ..., \alpha(p-1))$.*

The map ϕ also establishes the first isomorphism since it is a ring isomorphism. Hence, we have $R \cong Z_p[v]/\langle v\rangle \oplus Z_p[v]/\langle v-1\rangle \oplus ... \oplus Z_p[v]/\langle v-(p-1)\rangle \cong Z_p^p$. Since Z_p is a field then its ideals are trivial ones then the ideals of Z_p^p consist of the product of the trivial ones. Therefore R is a principal ideal ring.

Lemma 2.2 *The number of ideals of R is 2^p.*

The ideals of R are when ordered by the set inclusion, they do not give a single ordered chain. Such rings are called non chain rings.

If $r \in R$, then, the Hamming weight of a denoted by $w(a)$ is equal to zero if $a = 0$, and one otherwise. Further, if $a = (a_1, a_2, \ldots, a_n) \in R^n$, then the Hamming weight of a is defined by $w(a) = \sum_{i=1}^{n} w(a_i)$. Hamming distance (d) between two elements is the Hamming weight of their difference. It is a well known fact that Hamming distance is a metric on Z_p^n [9].

Lemma 2.3 *Let $I = \langle\alpha\rangle$ where $\alpha = a_0 + a_1 v + ... + a_{p-1}v^{p-1} \in R$.*

1. *If $\sum_{i=0}^{p-1} w(\alpha(i)) = p$, then α is a unit in R.*
2. *$|I| = p^{\sum_{i=0}^{p-1} w(\alpha(i))}$.*

Note that by applying the Chinese Remainder Theorem the inverse map of ϕ exists. Next, we can define a Gray weight w_G on R such that: Let $\alpha = a_0 + a_1 v + ... + a_{p-1}v^{p-1} \in R$.

$$w_G(\alpha) = w(\phi(\alpha)). \tag{1}$$

The Gray distance between any two elements α and β of R is defined by $d_G(\alpha, \beta) = w(\phi(\alpha) - \phi(\beta))$ which is a linear distance preserving map from (R^n, d_G) to (Z_p^{pn}, d).

For instance, let $p = 7$, if $\alpha = 4v+v^2$ and $\beta = 4v+4v^3+5v^4+v^5$, then $w_G(\alpha) = w(\phi(4v+v^2)) = w(0,5,5,0,4,3,4) = 5$ and $w_G(\beta) = w(\phi(4v+4v^3+5v^4+v^5)) = w(0,0,5,5,0,1,3) = 4$ hence $d_G(\alpha,\beta) = w(\phi(\alpha)-\phi(\beta)) = w((0,5,5,0,4,3,4)-(0,0,5,5,0,1,3)) = w((0,5,0,2,4,2,1)) = 5$.

Definition 2.4 *Let $a = (a_0, a_1, .., a_{p-1}) \in Z_p^p$. Then, $supp(a) = \{i | a_i \neq 0\} \subseteq \{1, 2, ..., p\}$.*

We can easily check that:

- Let $\alpha, \beta \in R$. If $supp(\phi(\alpha)) = supp(\phi(\beta))$ then $w_G(\alpha) = w_G(\beta)$.
- Let $\langle\alpha\rangle$ and $\langle\beta\rangle$ be two ideals in R. $supp(\phi(\alpha)) = supp(\phi(\beta))$ if and only if $\langle\alpha\rangle = \langle\beta\rangle$.

Theorem 2.5 *Let $I = \langle\alpha_1, \alpha_2, \ldots, \alpha_s\rangle$ be a finitely generated ideal of R. Then, $I = \langle\beta\rangle$ for some $\beta \in R$ where $supp(\phi(\beta)) = \bigcup_{i=1}^{s} supp(\phi(\alpha_i))$.*

Lemma 2.6 *Let $\alpha \in R$. The followings are true:*
i) If $supp(\phi(\alpha)) = \{1, ..., p-1\}$, then α is a unit. Thus, there are $(p-1)^p$ units in R.
ii) Let $I = \langle\alpha\rangle$ be an ideal of R. If $|\ supp(\phi(\alpha))| = p-1$, then I is maximal. Hence, again there are p maximal ideals in R.

3 Linear Codes over R

Since R is not a chain ring i.e the ideals do not form a single chain with respect to set inclusion, there is no straightforward way of expressing the generating matrix. For instance in [11] and in [14] some special definitions or cases are adapted in order to express linear codes by generating sets. Here, we make use of the map ϕ as above and by considering the image of the code we define its generator matrix and obtain the related results.

Theorem 3.1 *Let $\{g_1, g_2, \ldots, g_k\} \subset R^n$ be a generating set of elements of a linear code C over R of length n where $g_i = (g_{i1}, g_{i2}, \ldots, g_{in})$. Then, the matrix*

$$\phi(G) = \begin{bmatrix} \phi(g_{11}) & \phi(g_{12}) & \cdots & \phi(g_{1n}) \\ \phi(vg_{11}) & \phi(vg_{12}) & \cdots & \phi(vg_{1n}) \\ \phi(v^2g_{11}) & \phi(v^2g_{12}) & \cdots & \phi(v^2g_{1n}) \\ \vdots & \vdots & \vdots & \\ \phi(v^{p-1}g_{11}) & \phi(v^{p-1}g_{12}) & \cdots & \phi(v^{p-1}g_{1n} \\ \vdots & \vdots & \vdots & \\ \phi(g_{k1}) & \phi(g_{k2}) & \cdots & \phi(g_{kn}) \\ \phi(vg_{k1}) & \phi(vg_{k2}) & \cdots & \phi(vg_{kn}) \\ \phi(v^2g_{k1}) & \phi(v^2g_{k2}) & \cdots & \phi(v^2g_{kn}) \\ \vdots & \vdots & \vdots & \\ \phi(v^{p-1}g_{k1}) & \phi(v^{p-1}g_{k2}) & \cdots & \phi(v^{p-1}g_{kn} \end{bmatrix}.$$

generates $\phi(C)$.

Let $\alpha = g_0 + g_1v + ... + g_{p-1}v^{p-1} \in R$ and $\alpha(i) = g_0 + g_1i + ... + g_{p-1}i^{p-1}(mod\ p)$ for $1 \leq i \leq p-1$. Then, $\phi(\alpha) = (g_0 + g_1i + ... + g_{p-1}i^{p-1}(mod\ p)) = (\alpha(0), \alpha(1), ..., \alpha(p-1))$.

By applying some row operations, the generator matrix becomes row equivalent to a block matrix whose blocks are $(G_{ij})_{k\times n}$ where

$$G_{ij} = \begin{bmatrix} \alpha(0) & 0 & \vdots & 0 & \\ 0 & \alpha(1) & \vdots & 0 & \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \vdots & \alpha(p-1) & \end{bmatrix}.$$

R is not a chain ring so minimal independent sets that generate a submodule (linear code) over R can not be defined directly. We adopt a similar approach as in [11] as follows:

Definition 3.2 *A set* $\{g_1, g_2, \ldots, g_k\} \subset R^n$ *is called a minimal independent generating set for a code* C, *if*

$$\{\phi(g_1), \phi(vg_1), \ldots, \phi(v^{p-1}g_1), \phi(g_2), \phi(vg_2), \ldots, \phi(v^{p-1}g_2), \ldots, \phi(g_k), \phi(vg_k), \ldots, \phi(v^{p-1}g_k)\} \subset Z_p^{pn}$$

is a Z_p*-linearly independent set.*

Lemma 3.3 *If* $C = \langle\{g_1, g_2, \ldots, g_k\}\rangle$ *where the set* $\{g_1, g_2, \ldots, g_k\} \subset R^n$ *is a minimal independent generating set, then* $|C| = p^{pk}$.

Let $g = [g_1, g_2, \ldots, g_n], h = [h_1, h_2, \ldots, h_n] \in R^n$, $g_i = g_{i1} + g_{i2}v + \ldots + g_{ip}v^{p-1}$,$h_i = h_{i1} + h_{i2}v + \ldots + h_{ip}v^{p-1}$, $g_i(j) = g_{i1} + g_{i2}j + \ldots + g_{ip}j^{p-1}(\mod p)$ and $h_i(j) = h_{i1} +$

$h_{i2}j + \ldots + h_{ip}j^{p-1} (\mod p)$ for $1 \leq j \leq p$. An inner product between the elements of g and h is defined as follows:

$$\langle g, h \rangle_\phi = \sum_{i=1}^{n} \sum_{j=1}^{p-1} (g_i(j) h_i(j)) .$$

Let C be a linear code of length n over R. Then the dual code of C is defined by

$$C^\perp = \{h \in R^n | \langle g, h \rangle_\phi = 0 \text{ for all } g \in C\}. \tag{2}$$

Lemma 3.4 $\phi(C)^\perp = \phi(C^\perp)$.

4 Cyclic Codes over R

The family of cyclic codes is an important class of linear codes and have generated great interest in coding theory. Moreover, since they can be described as ideals in same polynomial rings, they have a rich algebraic structure hence they are easy to implement in engineering. In this section we study the algebraic structure of cyclic codes over the ring R. We also conclude by presenting the structure of dual codes.

Definition 4.1 *A linear code of length n over R is called a cyclic code if every right cyclic shift of a codeword in C also falls in C. In other word a linear code of length n is cyclic if it is invariant under automorphism σ which is $\sigma(c_0, c_1, \ldots, c_{n-1}) = (c_{n-1}, c_0, \ldots, c_{n-2})$.*

As usual if we can identify a codeword $c = (c_0, c_1, \ldots, c_{n-1})$ in a cyclic code with a polynomial $c(x) = c_0 + c_1 x + \cdots + c_{n-1} x^{n-1}$, then we obtain a subset of $R[x]$. The cyclic property above means that if $c(x) \in C$ so is $xc(x)$ modulo $x^n - 1$. Since a cyclic code of length n is linear and is closed under multiplication by x modulo $x^n - 1$, it is then clear that a cyclic code C over R is an ideal in the quotient ring $R[x]/\langle x^n - 1 \rangle$.

Let $R_n = R[x]/\langle x^n - 1 \rangle$. Since $R \cong Z_p^p$, then

$$R[x]/\langle x^n - 1 \rangle \cong Z_p[x]/\langle x^n - 1 \rangle \times Z_p[x]/\langle x^n - 1 \rangle \times \ldots \times Z_p[x]/\langle x^n - 1 \rangle.$$

Let

$$L_n = Z_p[x]/\langle x^n - 1 \rangle \times Z_p[x]/\langle x^n - 1 \rangle \times \ldots \times Z_p[x]/\langle x^n - 1 \rangle.$$

Now we can extend the map ϕ to get an isomorphism between the rings R_n and L_n in a natural way. First, for $1 \leq i \leq p$, we define a projection map

$$\pi_i : Z_p^p \to Z_p,$$

such that $\pi_i((a_1, a_2, \ldots a_p)) = a_i$.

Next, we define

$$\begin{aligned}\phi :& R_n \to L_n\\ &\phi(\sum_{i=0}^{n} a_i x^i) = (\sum_{i=0}^{n} \pi_1(\phi(a_i))x^i, \sum_{i=0}^{n} \pi_2(\phi(a_i))x^i, \ldots, \sum_{i=0}^{n} \pi_p(\phi(a_i))x^i).\end{aligned}$$

For instance, let $p = 5$ and if $f(x) = (1 + v^2)x^3 + (1 + 3v + v^3)x^2 + (3v^2 + v^4)x + 1$ in R_4, then, $\phi(f(x)) = (x^3 + x^2 + 1, 2x^3 + 4x + 2, 3x + 1, 2x^2 + 3x + 1, 2x^3 + 2x^2 + 4x + 1)$.

Since we know the structure of $Z_p[x]/\langle x^n - 1\rangle$ we can get the structure of $R_n = R[x]/\langle x^n - 1\rangle$. So, R_n is a principal ideal ring. Further, we can determine the generator of ideals as follows: Let $I = \langle f_1(x), f_2(x), \ldots, f_s(x)\rangle$ be a finitely generated ideal of R_n where $f_i(x) = \sum_{j=0}^{n} f_{ij}x^j$ and $g_i = gcd_{1\leq j\leq s}(\pi_i(\phi(f_j))), x^n - 1)$ for $i = 1, 2, \ldots, p$. Hence, $I = \langle g(x)\rangle$ where $g(x) = \phi^{-1}((g_1(x), g_2(x), \ldots, g_p(x)))$.

Lemma 4.2 *Let $C = \langle g(x)\rangle$ be a cyclic code of length n over R and $\phi(g(x)) = (g_1, g_2, \ldots, g_p)$ with* $\deg(gcd(g_i, x^n - 1)) = n - k_i$ *for $1 \leq i \leq p$, , Then, $|C| = p^{\sum_{i=1}^{p} k_i}$.*

Now we present the structure of the dual of a cyclic code. Suppose $C = \langle g(x)\rangle$ is a cyclic code of length n over R and $g_i = \pi_i(\phi(g(x)))$, then $\phi(C) = J = \langle(g_1(x), g_2(x), \ldots g_p(x))\rangle$. The dual of J is the cyclic code

$$J^\perp = \langle(h_{1_R}(x), h_{2_R}(x), \ldots, h_{p_R}(x))\rangle,$$

where $h_i(x) = (x^n - 1)/(gcd(x^n - 1, g_i)$. Let $h_{i_R}(x)$ be the reciprocal polynomial of $h_i(x)$. Hence, $C^\perp = \langle\phi^{-1}(h_{1_R}(x), h_{2_R}(x), \ldots, h_{p_R}(x))\rangle$.

5 Conclusion

We have extended the work done in [1] nd introduced a non chain ring. The structure of this ring and codes both linear and cyclic are studied. As a non chain ring some other properties of codes defined on this ring can be of interest to the researchers.

References

[1] A. BAYRAM AND I. SIAP, *Structure of Codes over the Ring $Z_3[v]/\langle v^3 - v\rangle$*, Applicable Algebra in Engineering, Communication and Computing (Accepted 2013.)

[2] T. ABUALRUB AND I. SIAP, *On the construction of cyclic codes over the ring $Z_2 + uZ_2$*, WSEAS Trans. on Math. **Issue 6, Vol. 5** (June 2006) 750-756.

[3] M. Al-Ashker and M. Hamoudeh, *Cyclic codes over $Z_2+uZ_2+u^2Z_2+...+u^{k-1}Z_2$*, Turk. J. Math. **Vol. 35** (2011) 737-749.

[4] T. Abualrub and I. Siap, *Cyclic codes over the rings Z_2+uZ_2 and $Z_2+uZ_2+u^2Z_2$*, Des. Codes Crypt. **Vol. 42, No. 3** (March 2007) 273-287.

[5] K. Betsumiya and M. Harada, *Optimal self-dual codes over $F_2 \times F_2$, with respect to the Hamming weight*, IEEE Trans. Inf. Theory **Vol. 50, No. 2,** (2004) 356-358.

[6] A. Bonnecaze and P. Udaya, *Cyclic codes and self-dual codes over F_2+uF_2,* IEEE Trans. Inf. Theory. **Vol.45, No.4,** (1999) 1250-1255.

[7] S.T. Dougherty, B. Yildiz and S. Karadeniz, *Codes over R_k, Gray maps and their binary images*, Finite Fields and Their Appl. **I Vol. 17** (2011) 205-219.

[8] A. R. Hammons, Jr. P. V. Kumar, J. A. Calderbank, N. J. A.Sloane and P. Sole, *The Z-linearity of Kerdock, Preparata, Goethals, and related codes*, IEEE Trans. Inf. Theory. **Vol.40, No.2** (1994) 301-319.

[9] F.J. MacWilliams and N. J. A.Sloane, *The theory of error correcting codes*, North-Holland, Amsterdam, The Netherlands, 1977.

[10] M. Ozen and I. Siap, *Linear codes over $F_q[u]/(u^s)$ with respect to the Rosenbloom–Tsfasman metric*, Des. Codes Crypt. **Vol. 38** (2006) 17-29.

[11] Y. H. Park, *Modular independence and generator matrices for codes over Z_m*, Des. Codes Crypt. **Vol. 50, No 2** (2009) 147-162.

[12] J.F. Qian, L.N. Zhang and Shi-Xin Zhu, *Constacyclic and cyclic codes over $F_2+uF_2+u^2F_2$*, IEICE Trans. Fundamentals. **Vol.E89–A, No.6** (June 2006) 1863-1865.

[13] B. Yildiz and S. Karadeniz, *Linear codes over $F_2+uF_2+vF_2+uvF_2$*, Des. Codes Crypt. **Vol. 54** (2010) 61-81.

[14] B. Yildiz and S. Karadeniz, *Cyclic codes over $F_2+uF_2+vF_2+uvF_2$*, Des. Codes Crypt. **Vol. 58, No.3** (2011) 221-234.

[15] Shi-Xin Zhu, Y. Wang and Min-Jia Shi, *Cyclic codes over F_2+vF_2*, ISIT'09 Proceedings of the 2009 IEEE International Conference on Symposium on Information Theory **Vol.3** (2009) 1719-1722.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Soft Path Connectedness on Soft Topological Spaces

S. Bayramov[1], **C. Gunduz**[2] **and A.Erdem**[2]

[1] *Department of Mathematics, Kafkas University, Kars, 36100-Turkey*

[2] *Department of Mathematics, Kocaeli University, Kocaeli, 41380-Turkey*

emails: baysadi@gmail.com, carasgunduz@gmail.com, erdem.arzu@gmail.com

Abstract

There are some theories as soft topological space and some related concepts such as soft interior, soft closed, soft subspace, soft separation axioms in [8] and soft connectedness, soft locally connectedness in [7]. In this paper, we define soft path connectedness on soft topological space and continue investigating the properties of soft path connectedness which is fundamental result for further research on soft topology.

Key words: Soft topological space, soft path connectedness.

1 Introduction

Since there was no mathematical concept to solve complicated problems in the economics, engineering,and environmental areas, a soft set theory was firstly introduced by Molodtsov [6] to deal with the various kinds of uncertainties in these problems. Many researchers have contributed towards the soft set theory and its applications in various fields, increasingly [1, 2, 5, 9, 10]. Recently the notion of soft topological spaces was studied by Shabir and Naz [8]. They also introduced the concepts of soft open sets, soft closed sets, soft interior, soft closure and soft separation axioms. In [7], soft connectedness, soft locally connectedness was investigated.

In this paper, we introduce some new concepts in soft topological spaces such as soft path connectedness and examine some properties and relations of this concept.

2 Preliminaries

We now recall some definitions.

Definition 2.1. ([6]) *Let X be an initial universe and E be a set of parameters. The power set of X is denoted by $P(X)$ and A is a non-empty subset of E .A pair (F,A) is called a soft set over X, where F is a mapping given by $F : A \to P(X)$. In other words, a soft set over X is a parameterized family of subsets of the universe X. For $e \in A$, $F(e)$ can be considered as the set of e-approximate elements of the soft set (F,A). Clearly, a soft set is not a set.*

Definition 2.2. ([6]) *The intersection of two soft sets (F,A) and (G,B) over X, is the soft set (H,C), where $C = A \cap B$, and $\forall c \in C$, $H(e) = F(e) \cap G(e)$. We write $(F,A) \cap (G,B) = (H,C)$.*

Definition 2.3. ([6]) *A soft set (F,A) over X is called a null soft set, denoted by Φ, if $\forall e \in A$, $F(e) = \emptyset$.*

Definition 2.4. ([6]) *A soft set (F,A) over X is called an absolute soft set, denoted by $\widetilde{A}$, if $\forall e \in A$, $F(e) = X$.*

Definition 2.5. ([6]) *The union of two soft sets (F,A) and (G,B) over X, is the soft set (H,C), where $C = A \cup B$, and and $\forall c \in C$*

$$H(e) = \begin{cases} F(e), & \text{if } e \in A - B, \\ G(e), & \text{if } e \in B - A, \\ F(e) \cup G(e), & \text{if } e \in A \cap B. \end{cases}$$

We write $(F,A) \cup (G,B) = (H,C)$.

Definition 2.6. ([5]) *Let (F,A) and (G,B) be two soft sets over X. Then (F,A) is a subset of (G,B), denoted by $(F,A) \tilde{\subset} (G,B)$ if $A \subset B$ and for all $e \in E$, $F(e) \subset G(e)$.*

Definition 2.7. ([3]) *A soft set (F,E) is called a soft point, denoted by (x_e, E), if for each element $e \in E$, $F(e) = \{x\}$ and $F(e') = \emptyset$ for all element $e' \in E - \{e\}$.*

Definition 2.8. ([8]) *Let τ be be the collection of soft sets over X, then τ is called a soft topology on X if τ satisfies the following axioms:*
(i) $\Phi, \widetilde{X}$ belong to τ.
(ii) The union of any number of soft sets in τ belongs to τ.
(iii) The intersection of any number of soft sets in τ belongs to τ.
The triplet (X,τ,E) is called a soft topological space over X. The members of τ are said to be soft open in X.

Definition 2.9. *([4])A soft set (G,E) in a soft topological space (X,τ,E) is called a soft neighborhood of $x\in X$ if there exists a soft open set (F,E) such that $x\in(F,E)\,\widetilde{\subset}\,(G,E)$.*

Definition 2.10. ([10]) *Let (X,τ,E) and (Y,τ',E) be two soft topological spaces, $f:(X,\tau,E)\to(Y,\tau',E)$ be a mapping. For each neighborhood (H,E) of $\left(f(x)_e,E\right)$, if there exists a soft neighborhood (F,E) of (x_e,E) such that $f((F,E))\,\widetilde{\subset}\,(H,E)$, the f is called a soft continuous mapping at (x_e,E).*

Definition 2.11. ([7]) *Let (X,τ,E) be a soft topological space over X.A soft separation of $\widetilde{X}$ is a pair (F,E) and (G,E) of no-null soft open sets over X such that*

$$\widetilde{X}=(F,E)\cup(G,E)\,;\;\;(F,E)\cap(G,E)=\Phi$$

Definition 2.12. ([7]) *A soft topological space (X,τ,E) is said to be soft connected if there does not exist a soft separation of $\widetilde{X}$.*

Definition 2.13. *([7]) Let $\{(X_s,\tau_s,E_s)\}_{s\in S}$ be a family of soft topological spaces and define $B=\left\{\prod\limits_{s\in S}(F_s,E_s):(F_s,E_s)\in\tau_s\right\}$ and τ as the collection of all arbitrary union of elements of B and τ is a soft topology over $\prod\limits_{s\in S}(X_s,\tau_s,E_s)$.*

Unless otherwise stated $E=\mathbb{N}\cup\{0\}$ will be assumed to be a set of parameters and the set of rational numbers on the closed interval $I=[0,1]$ will be considered as $\{0,1,r_3,r_4,r_5,\ldots\}$. If $e\in E, r_e\in\mathbb{Q}\cap I$ and for all $\varepsilon>0$ we define the soft set $F_\varepsilon:E\to P(I)$ as $F_\varepsilon(e)=(r_e-\varepsilon,r_e+\varepsilon)$. Then the family $\mathcal{B}=\{(F_\varepsilon,E)\}_{\varepsilon>0}$ is a soft base of soft toplogy on I. Then τ_I is called a soft toplogy generated by $\mathcal{B}$.

Definition 2.14. *A soft topological space (I,τ_I,E) is called a unit soft interval.*

Definition 2.15. *Let (I,τ_I,E) be a unit soft interval and (X,τ,E') be a soft topological space. A soft path is a soft continous map $(f,\varphi):(I,\tau_I,E)\to(X,\tau,E')$ The soft sets $\left\{f(0)_{\varphi(e)}\right\}_{e\in E}$ and $\left\{f(1)_{\varphi(e)}\right\}_{e\in E}$ are said to be the initial and final of the soft path (f,φ) where $f:I\to X,\varphi:E\to E'$. It's clear that for each $e\in E$, the map $f:(I,\tau_I)\to\left(X,\tau_{\varphi(e)}\right)$ is a path from $f(0)_{\varphi(e)}$ to $f(1)_{\varphi(e)}$. Hence every soft path can be considered as a parametrized family on the soft topological space (X,τ,E').*

Definition 2.16. *Let (I,τ_I,E) be a unit soft interval and (X,τ,E') be a soft topological space. (X,τ,E') is said to be a soft path connected space if for each soft points $\left(x_{e'_1},E'\right),\left(y_{e'_2},E'\right)$, there exists a soft path $(f,\varphi):(I,\tau_I,E)\to(X,\tau,E')$ such that*

$$\varphi(e_1)=e'_1,\varphi(e_2)=e'_2,f(0)=x,f(1)=y$$

3 The Main Results

The following proposition shows the relation between the soft path connected topological space and path connected topological space.

Proposition 3.1. *If (X, τ, E') is a soft path connected topological space, then $(X, \tau_{e'})$ is a path connected topological space for each $e' \in E'$*

Proof. Let $x, y \in (X, \tau_{e'})$ be arbitrary points; then $(x_{e'}, E'), (y_{e'}, E') \in (X, \tau, E')$ are soft points. Since (X, τ, E') is a soft path connected topological space, there exists a soft path $(f, \varphi) : (I, \tau_I, E) \to (X, \tau, E')$ such that

$$\varphi(e) = e', f(0) = x, f(1) = y.$$

So for each $e' \in E'$, we now have a path from x to y defined by

$$f_e : (I, (\tau_I)_e) \to (X, \tau_{e'})$$

This implies that $(X, \tau_{e'}), \forall e' \in E'$ is a path connected topological space. □

Proposition 3.2. *The soft unit interval (I, τ_I, E) is a soft path connected topological space.*

Proof. Since $(1_I, 1_E) : (I, \tau_I, E) \to (I, \tau_I, E)$ is a soft continous map, this is trivial. □

The converse of Proposition 3.1 is is not always true, as shown in the next example.

Example 3.3. *Let (X, τ_1) and (Y, τ_2) be two path connected and disjoint topological spaces. Then for $E_1 = \{e_1\}, E_2 = \{e_2\}$ we can construct the following soft topological spaces*

$$\{F_U : E_1 \to P(X) : F_U(e_1) = U, U \in \tau_1\}$$
$$\{F_V : E_2 \to P(Y) : F_V(e_2) = V, V \in \tau_2\}$$

and consider the soft toplogical space $(X \oplus Y, \tau_1 \oplus \tau_2, E_1 \cup E_2)$. For any $c_1 \in E_1 \cup E_2$, $\left(X \oplus Y, (\tau_1 \oplus \tau_2)_{e_1}\right) = (X, \tau_1)$ and $\left(X \oplus Y, (\tau_1 \oplus \tau_2)_{e_2}\right) = (Y, \tau_2)$ are path connected topological spaces but $(X \oplus Y, \tau_1 \oplus \tau_2, E_1 \cup E_2)$ is not a soft path connected topological space.

Theorem 3.4. *The image under a soft continuous map of a soft path connected topological space is soft path connected.*

Proof. Suppose (X, τ, E') is a soft path connected topological space, (Y, τ', E'') is a soft topological space and $(g, \psi) : (X, \tau, E') \to (Y, \tau', E'')$ is a soft continuous map. Let $\left(y_{e''_1}, E''\right)$ and $\left(\overline{y}_{\overline{e}''_1}, E''\right)$ be soft points of the image $(g, \psi)(X, \tau, E')$; then there exist soft points $\left(x_{e'_1}, E'\right)$, $\left(\overline{x}_{\overline{e}'_1}, E'\right) \in (X, \tau, E')$ so that

$$\psi(e'_1) = e''_1, \psi(\overline{e}'_1) = \overline{e}''_1, g(x) = y, g(\overline{x}) = \overline{y}.$$

Since (X,τ,E') is a soft path connected topological space, we have a soft path $(f,\varphi):(I,\tau_I,E)\to(X,\tau,E')$ such that

$$\varphi(e_1)=e_1',\varphi(\overline{e}_1)=\overline{e}_1',f(0)=x,f(1)=\overline{x}.$$

Hence $(g,\psi)\circ(f,\varphi):(I,\tau_I,E)\to(Y,\tau',E'')$ is a soft path from soft point $\left(y_{e_1''},E''\right)$ to soft point $\left(\overline{y}_{\overline{e}_1''},E''\right)$. □

Theorem 3.5. *The soft unit interval (I,τ_I,E) is a soft connected topological space.*

Proof. We argue by contradiction. Suppose that the soft unit interval (I,τ_I,E) is not a soft connected topological space. Then

$$(F,E)\cup(G,E)=(I,\tau_I,E)$$

where (F,E) and (G,E) are nonempty, disjoint soft sets. For each $e\in E$, choose nonempty sets $F(e)$, $G(e)\in(\tau_I)_e$ and then

$$F(e)\cup G(e)=I,F(e)\cap G(e)=\emptyset.$$

It follows that $(I,(\tau_I)_e)$ is not a connected topological space, which is a contradiction. □

Theorem 3.6. *Any soft path connected topological space is a soft connected topological space.*

Proof. Assume for contradiction that (X,τ,E') is soft path connected topological space but not a soft connected topological space. Then there exist nonempty, disjoint soft sets (F,E') and (G,E') so that

$$\left(F,E'\right)\cup\left(G,E'\right)=\widetilde{X}.$$

Since (X,τ,E') is soft path connected topological space, for each soft points $\left(x_{e_1'},E'\right)\in(F,E'),\left(y_{e_2'},E'\right)\in(G,E')$ there exists a soft path $(f,\varphi):(I,\tau_I,E)\to(X,\tau,E')$ such that

$$\varphi(e_1)=e_1',\varphi(e_2)=e_2',f(0)=x,f(1)=y.$$

By Theorem 3.5, (I,τ_I,E) is a soft connected topological space and Theorem 2.8 in [7] implies that $(f,\varphi)(I,\tau_I,E)$ is a soft connected topological space. Suppose that

$$\begin{array}{rcl}\left(F_1,E'\right)&=&\left(F,E'\right)\cap(f,\varphi)(I,\tau_I,E)\\ \left(G_1,E'\right)&=&\left(G,E'\right)\cap(f,\varphi)(I,\tau_I,E)\end{array}$$

are two disjoint soft sets on the soft space $(f,\varphi)(I,\tau_I,E)$ such that

$$(f,\varphi)(I,\tau_I,E)=\left(F_1,E'\right)\cup\left(G_1,E'\right).$$

This contradicts the fact that $(f,\varphi)(I,\tau_I,E)$ is a soft connected topological space. □

Theorem 3.7. *The topological product of soft topological spaces is a soft path connected topological space if and only if each soft topological space is a soft path connected topological space.*

Proof. Let $\{(X_s, \tau_s, E_s)\}_{s\in S}$ be a family of soft topological spaces and $\prod\limits_{s\in S}(X_s, \tau_s, E_s)$ be a soft path connected topological space. Since the projections $(p_s, q_s) : \prod\limits_{s\in S}(X_s, \tau_s, E_s) \to (X_s, \tau_s, E_s)$ are soft continuous and surjective, for each $s \in S$, $\{(X_s, \tau_s, E_s)\}_s$ is a soft path connected topological space.
Conversely, Let $\{(X_s, \tau_s, E_s)\}_{s\in S}$ be a family of soft path connected topological spaces and $\left\{\left((x_s)_{e'_s}, E_s\right)\right\}_{s\in S}, \left\{\left((y_s)_{e'_s}, E_s\right)\right\}_{s\in S}$ be soft points of $\prod\limits_{s\in S}(X_s, \tau_s, E_s)$. Then for each $s \in S$, $\left((x_s)_{e'_s}, E_s\right)$ and $\left((y_s)_{e'_s}, E_s\right)$ belong to (X_s, τ_s, E_s). Since (X_s, τ_s, E_s) is a soft path connected topological space, we have a soft path $(f_s, \varphi_s) : (I, \tau_I, E) \to (X_s, \tau_s, E_s)$ such that

$$\varphi_s(e) = e_s, \varphi_s(\overline{e}) = \overline{e}_s, f_s(0) = x_s, f_s(1) = y_s.$$

If we now take $\varphi : E \to \prod\limits_{s\in S} E_s, f : I \to \prod\limits_{s\in S} X_s$ defined by $\varphi(e) = \{\varphi_s(e)\}_{s\in S}, f(x) = \{f_s(x)\}_{s\in S}$, then $(f, \varphi) : (I, \tau_I, E) \to \prod\limits_{s\in S}(X_s, \tau_s, E_s)$ is a soft path so that

$$\begin{aligned}\varphi(e) &= \{\varphi_s(e)\}_{s\in S} = \{e_s\}_{s\in S}, \varphi(\overline{e}) = \{\varphi_s(\overline{e})\}_{s\in S} = \{\overline{e}_s\}_{s\in S},\\ f(0) &= \{f_s(0)\}_{s\in S} = \{x_s\}_{s\in S}, f(1) = \{f_s(1)\}_{s\in S} = \{y_s\}_{s\in S}.\end{aligned}$$

□

References

[1] Ahmad, B. and Kharal, A., *On fuzzy soft sets*, Advances in Fuzzy Systems, 2009(2009), 6 pages.

[2] Alkhazaleh, S. and Salleh, A.R. and Hassan, N., *Possibility fuzzy soft set*, Advances in Decision Sciences, 2011(2011), 18 pages.

[3] Bayramov, S. and Gunduz, (Aras) C. , *Soft locally compact and soft paracompact spaces*, Journal of Mathematics and System Science, (2013), accepted.

[4] Hussain, S. and Ahmad, B. , *Some properties of soft topological spaces*, Comput. Math. Appl., 62(2011), 4058-4067.

[5] Maji, P. K. and Biswas, R. and Roy, A. R. , *Fuzzy soft sets*, Journal of Fuzzy Mathematics, 9-3(2001), 589-602.

[6] Molodtsov, D., *Soft set theory-first results*, Computers & Mathematics with Applications, 37(1999), 19-31.

[7] Peyghan, E. and Samadi, B. and Tayebi, A., *On Soft Connectedness*, arXiv:1202.1668v1 [math.GN], 8 Feb 2012.

[8] Shabir, M. and Naz, M., *On soft topological spaces*, Computers and Mathematics with Applications, 61(2011), 1786-1799.

[9] Zhou, X. and Li, Q. and Guo, L. , *On generalized interval-valued fuzzy soft sets*, Journal of Applied Mathematics, 2012(2012), 18 pages.

[10] Zorlutuna, İ. and Akdag, M. and Min, W.K. and Atmaca, S., *Remarks on soft topological spaces*, Annals of Fuzzy Mathematics and Informatics, 3-2(2012), 171-185.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Problem Solving Environment for Gas Flow Simulation in Micro Structures on the Base of the Boltzmann Equation

I. I. Bazhenov[1], O. I. Dodulad[1,2], I. D. Ivanova[1], Yu. Yu. Kloss[1,2], V.V. Rjabchenkov[1], P. V. Shuvalov[2] and F.G. Tcheremissine[2,3]

[1] *, National Research Centre Kurchatov Institute*

[2] *, Moscow Institute of Physics and Technology*

[3] *, Dorodnicyn Computing Centre of RAS*

emails: , `dodulad@list.ru`, , , , ,

Abstract

The paper presents a description of the problem solving environment designed for simulation of gas flows in micro structures. The computational core of the problem solving environment is based on the direct finite-difference solution of the Boltzmann kinetic equation using the conservative projection method of collision integral calculation. To demonstrate the capabilities of the software and the present stage of its development some results of computing gas flows in the plane and 3D geometry are given. A simple gas, as well as a gas mixture, is considered.

Key words: problem solving environment, gas flow simulation, micro structures, Boltzmann equation

1 Introduction

At the present time, as technology development allows one to construct different micro and nano devices, the importance of understanding and predicting processes at the micro level has been increasing. One of the important areas of research is the study of non-equilibrium gas flows in microstructures. This area is difficult for theoretical study, and experimental study is often too expensive and time-consuming. Therefore, a promising research direction is numerical simulation. The traditional approach of gas flows simulation is based on the Navier-Stokes equations where the gas is treated as a continuous medium. This approach is applicable only in cases where the Knudsen number, i. e. the ratio of the molecular mean free path to the representative size of the system, is less than the order of 10^{-2}. However, in micro devices under normal conditions the mean free path is often comparable to the characteristic size of the flow, this leads to the inapplicability of the Navier-Stokes equations. For example, at the micro level there is an effect of the thermal transpiration where a temperature gradient applied to a surface causes a gas flow along this surface. Such effects do not appear at the macro level and not result from the solution of the Navier-Stokes equations.

A kinetic approach is required to describe the rarefied gas flows. The basis of the kinetic theory of gases is the Boltzmann equation. However, the complex structure and the nonlinearity of the Boltzmann equation cause difficulties in the development of not only analytical but also numerical methods for solving the equation.

A few decades ago a reliable method of the direct solution of the Boltzmann equation was developed [1]. Subsequently, the method has been improved and applied to the problems of highly non-equilibrium flows [3] as well as to the problems of slightly disturbed flows [2]. The method ensures the preservation of mass, momentum and energy conservation laws. It also turns the collision integral of the Maxwellian distribution function into zero and thus preserves the thermodynamic equilibrium state. The conservativeness of calculation of the collision integral is provided by a special projection of molecular velocities after collision at the nearest nodes of the velocity (or momentum) grid. Because of this procedure the method is called conservative projection method.

It is on the basis of the conservative projection method that the problem-solving environment (PSE) designed for simulation of rarefied gas flows was developed. Originally the PSE allowed one to carry out calculations for a one-component gas in areas which could be filled with a structured grid [4].

Then the PSE was improved, it became possible to carry out calculations of devices with complex geometries by means of unstructured grids [5]. Then the part of the PSE, which is responsible for the calculation of the collision integral, was improved; it became possible to simulate gas mixture flows [6] and gases with internal degrees of freedom on the basis of the simplified two-level model [7]. Most recently, the method of calculation of the collision integral for the gas mixture was extended [8], this was improved the efficiency of calculations of gas mixtures with disparate molecular masses. At present the integration into the PSE of the generalized Boltzmann equation solving method is taking place [7].

At the present time, there are also two other common approaches to modeling of rarefied gas flows. They are the direct Monte-Carlo simulation (DSMC), and the use of model equations, in which the complex collision integral is replaced by simpler relaxation forms. The first approach is used for the calculation of supersonic flows and, more recently, for micro-flows. However, for relatively slow subsonic processes, this method requires very time-consuming computation due to considerable statistical errors of $O(N^{-1/2}$. An example of a software environment based on the direct Monte-Carlo simulation is presented in [10]. The drawback of the second approach is that the reliability of the results obtained by the application of simplified models is unknown. A description of another solver designed for computation of rarefied gas flows which is also based essentially on the conservative projection method is given in [11].

2 General structure of problem solving environment

The problem solving environment was developed using open technologies. For parts of the PSE related to calculations C++ was chosen as a programming language because it allows one to combine a high performance with object-oriented programming. The last is important for the further prospective PSE development. PSE parts responsible for pre- and post-processing of data were written in a high-level programming language Python. The PSE allows one to perform calculations on a personal computer, including the use of graphics processors [12]. Complex two- and three-dimensional problems are computed on multi-processor clusters. The scheme of data flow from the input parameters provided by the user and the computational grid generation to the results visualization by means of specialized packages is presented in Fig. 1. In the center of the picture there is a solver that is a program that performs all necessary calculations for solving the Boltzmann equation. The calculation is an iterative process of the evolution of the distribution function $f(t, \boldsymbol{x}, \boldsymbol{\xi})$. The gas macroparameters are obtained by integrating the distribution function multiplied by the corresponding function of the velocity $\boldsymbol{\xi}$. For a simple gas the Boltzmann equation and the formulae for calculation of the density, velocity and temperature of the gas are the following:

$$\frac{\partial f(t,\boldsymbol{x},\boldsymbol{\xi})}{\partial t} + \boldsymbol{\xi}\frac{\partial f(t,\boldsymbol{x},\boldsymbol{\xi})}{\partial \boldsymbol{x}} = \int \left(f(\boldsymbol{\xi}_1')f(\boldsymbol{\xi}') - f(\boldsymbol{\xi}_1)f(\boldsymbol{\xi})\right) g\sigma(\theta, g)\, d^2\boldsymbol{n}\, d^3\boldsymbol{\xi},$$

$$n(t,\boldsymbol{x}) = \int f(t,\boldsymbol{x},\boldsymbol{\xi})d^3\boldsymbol{\xi},\ \boldsymbol{v}(t,\boldsymbol{x}) = \frac{1}{n}\int f(t,\boldsymbol{x},\boldsymbol{\xi})d^3\boldsymbol{\xi},\ T(t,\boldsymbol{x}) = \frac{m}{3kn}\int (\boldsymbol{\xi}-\boldsymbol{v})^2 f(t,\boldsymbol{x},\boldsymbol{\xi})d^3\boldsymbol{\xi}.$$

Since the method of solving the Boltzmann equation involves the splitting scheme into physical processes where the transport of molecules and collisions are calculated by turns, the solver itself consists of separate parts responsible for the transport equation and the calculation of the collision integral.

2.1 Molecular transport: advection equation solvers

The PSE includes two advection equation solvers: solver on structured grid (RectSolv) and solver on unstructured grids. In both solvers first and second order approximation schemes are implemented. The applied second order schemes are monotone and impervious

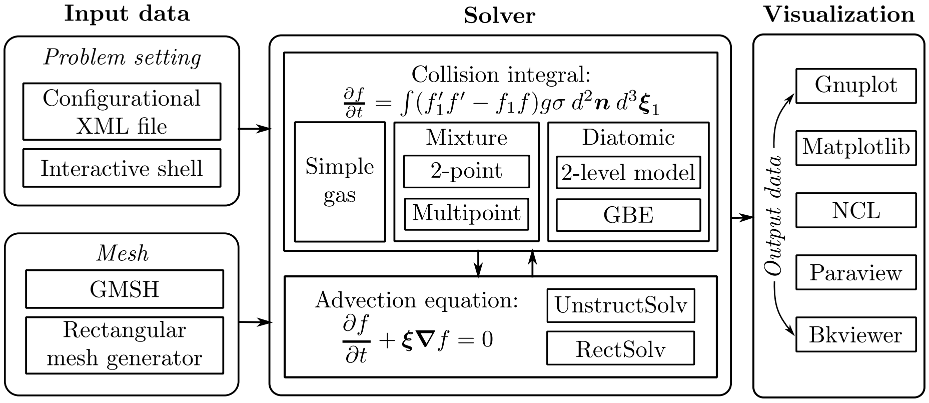


Figure 1: Scheme of the problem-solving environment.

to the appearance of oscillations. Structured grid solver is much easier and faster, and can be performed on the GPU, but its applicability is limited to only the problems where geometry allows the application of structured grids. Unstructured solver does not have such a drawback and it is a versatile tool suitable for simulation of complex devices. An open-source software package GMSH [13] is used for generation of unstructured grids. GMSH package allows one to fill the computational domain with the mesh satisfying the following two important requirements:

- The mesh should have a high quality. In the case of a tetrahedral mesh it means that each cell must be as possible close to a regular tetrahedron. This is due to the fact that the explicit numerical schemes in which the maximum allowable time step is limited by the Friedrichs – Courant – Lewy condition are used.
- The mesh should be more detailed in areas where the gas flow is of most interest and it should be less detailed where the parameters of gas vary slightly. This ensures optimal balance between performance and accuracy.

GMSH also takes into account the specifics of the problem and allows one to select the appropriate type of the unit cells. For example in three-dimensional problems where the computational domain can be obtained by extruding of two-dimensional field, unstructured tetrahedral mesh is not necessary. For such problems prismatic grids constructed by translations of unstructured triangular mesh for two-dimensional area are more suitable. The application of prismatic meshes also makes it easier to perform calculated domain decomposition and to achieve load balancing between processes as it will be discussed below.

Advection equation solvers also allow one to perform parallel calculations on multiprocessor cluster architectures. For this purpose the whole computational domain is divided into sub-areas (domains), each of which is provided by separate computing processor.

2.2 Collision integral solver

A key part of the PSE is a part of the solver responsible for the calculation of the collisions. According to the inbuilt method the collision integral is calculated by the quadrature formula that is conservative by mass, momentum, and energy. Due to the multidimensionality of the collision integral the Korobov grids are used to improve the accuracy of calculations. Overall calculation of the collision integral and the collision relaxation stage is performed according to the scheme of "continuous computation" [2] in which the distribution function varies continuously in the process of summation. The collision integral solver includes different potentials of intermolecular interaction. For example, the realistic Lennard-Jones potential can be applied.

The collision integral solver consists of five sub-parts: the part designed to simulate a simple gas, two parts for a mixture of gases using the two-point or the multi-point projection [8], respectively, and two parts for gas with internal degrees of freedom based on the two-level model and the generalized Boltzmann equation [9] respectively (the latter is currently under development).

The type of the simulated gas does not affect the transport process solvers, i. e., each of the sub-parts of the collision integral can be used with any of the advection equation solvers. The collision integral solver demands significant computational resources so its code has been carefully optimized by using of the processor vector SSE instructions.

2.3 Pre- and post-processing of data

The input parameters for the solver are specified in the configuration xml format file. The main settings in the configuration file are: the details of the computational grid in the physical space, the initial conditions specified as macro parameters or the distribution function, the boundary conditions, the type of simulated gas (a simple gas, a mixture of gases, a gas with internal degrees of freedom) and the molecule interaction potential, and the velocity/momentum grid characteristics in the case of a simple/mixture gas, respectively.

The auxiliary scripts written in Python programming language are used to generate the part of the configuration file that contains the information about the computational grid. For example, there is a script that converts GMSH file into the appropriate section of the configuration file. The remaining sections of the configuration file can be created in a text or xml editor or in a special GUI program. There is also a script used for parametric investigations that generates a large number of configuration files varying one of the parameters. The boundary conditions can be specified as follows: the conditions of specular reflection of molecules can be applied on the surfaces of symmetry of the problem, the conditions of diffuse reflection of molecules can be applied on the surface of contact with the solid walls (full or partial accommodation is possible), the strict conditions, for example, can be applied to simulate the flow of gas from the outside.

Different post-processors based on open source packages are used for the analysis of the results obtained. GnuPlot and Matplotlib library for Python are applied for preparation of camera ready plots and fields. The Paraview environment is convenient to operate with three-dimensional distributions of macro parameters. There is also our own product Bkviewer designed for rapid analysis of the output data.

It should be noted that the initial and boundary conditions input, the calculations and the results output are performed in dimensionless variables:

$$\boldsymbol{x}^* = \frac{\boldsymbol{x}}{\lambda},\ t^* = \frac{t}{\tau},\ \boldsymbol{\xi}^* = \frac{\boldsymbol{\xi}}{v_0},\ n^* = \frac{n}{n_0},\ \boldsymbol{v}^* = \frac{\boldsymbol{v}^*}{v_0},\ T^* = \frac{T}{T_0},\ f^* = \frac{f}{n_0 v_0^3},$$

where $\lambda = 1/\sqrt{2}\pi\sigma_0^2 n_0$ – the representative molecules mean free path, $\tau = \lambda/v_0$ – the mean free path time, v_0 – the representative velocity of the gas molecules, n_0, T_0 – the representative density and temperature of the gas. The transition to the dimensionless variables, firstly, facilitates the analysis of calculations, secondly, the criterion of similarity can be applied.

3 Examples of simulation

3.1 Poiseuille flow

Apart from fact that the Poiseuille flow problem is a classical problem of gas dynamics and can be used as a verification test, study of the Poiseuille flow through micro cavities and micro tubes has a practical interest in view of construction of vacuum devices, filter membranes and gas analyzers. In [14] and [15] the problem is studied by the DSMC method and solving of model equations, respectively. Here we present the results obtained by the developed PSE.

Let us consider a cylindrical tube with diameter $D = 4\lambda$ and length $L = 16\lambda$ (λ is the molecules mean free path), which connects two reservoirs. At the initial time there is a pressure difference at the ends of the tube so that the initial pressure ratio at the reservoirs is equal to $p_1/p_2 = 1.1$. An unstructured mesh built in the physical space is shown in Fig.2. The condition of the diffuse molecule reflection is set at the reservoirs and tube walls.

The pressure difference at the ends of the tube makes the Poiseuille flow in it. Fig. 3 shows the distribution of the gas velocity magnitude v after the flow between the reservoirs has established. In Fig. 4 the longitudinal velocity v_x versus the lateral coordinate y is shown. The solid line corresponds to the results of simulation, the dashed line is a quadratic approximation $u(x) = C - \Delta p x^2/2\mu$. The difference between the flow in the rarefied regime and the one in the continuous regime is primarily that the gas velocity at the walls of the tube is not equal to zero. In Fig. 5 the time evolution of the ratio of pressures in the reservoirs $p_1(t)/p_2(t)$ is given.

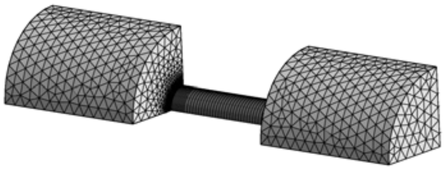


Figure 2: Unstructured mesh built in physical space.

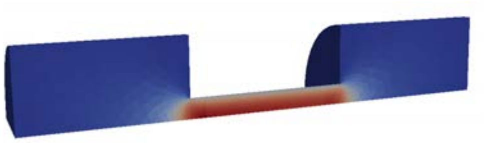


Figure 3: Distribution of gas velocity magnitude v.

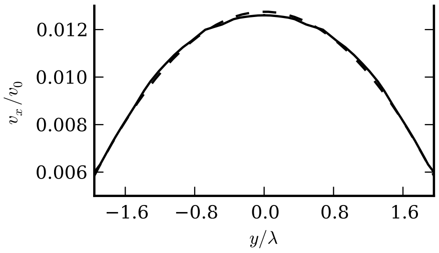


Figure 4: Longitudinal velocity v_x versus lateral coordinate$y$.

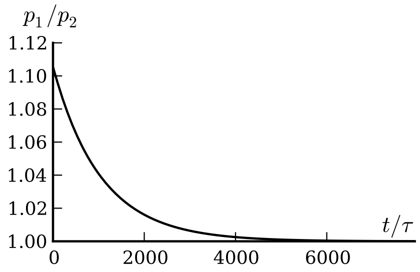


Figure 5: Pressure ratio versus time.

3.2 Shock wave – Boundary Layer Interaction inside a Micro Channel

Study of penetration of a shock wave into a boundary layer and its interaction with a channel wall is one of fundamental problems of gas dynamics. The problem is particularly important for analyses of flows with shock waves in channels of size comparable with the molecular mean free path [16]. Often there are problems related with experimental analyses of supersonic flows in micro channels, so using numerical modeling seems to be a promising approach.

For this problem the structured mesh solver of the advection equation is used. The calculations are performed in a rectangular channel. The channel has a width H and a length L. At the walls of the channel the diffuse reflection conditions are set. The temperature of the walls is constant and equal to the temperature of the gas before the shock wave. Maxwellian distributions with the parameters obtained from Rankin-Hugoniot relations are set as the boundary conditions at the ends of the channel. At the initial instant a non-disturbed shock wave structure calculated in advance in a one-dimensional problem is placed in the channel. After some period of time, a boundary layer structure is formed. The calculations are performed in the shock wave front coordinate system. The hard sphere intermolecular interaction model is used.

Figs. 6(a) and 6(b) present fields of density and temperature for the shock wave Mach

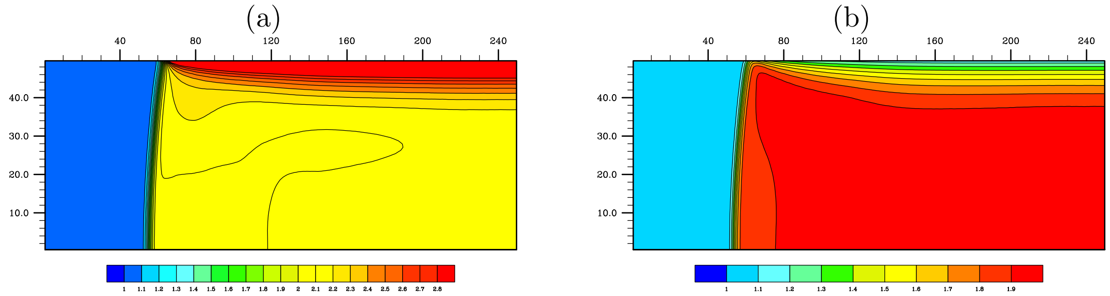


Figure 6: Field of (a) density and (b) temperature for $\mathrm{Kn} = 0.01$, $\mathrm{M} = 2$, $t = 25$

number $\mathrm{M} = 2$ and the Knudsen number $\mathrm{Kn} = \lambda/H$ at the moment of time $t = 25$. In Figs the bend of the shock wave front is observed. The hot gas behind the front of the SW is slowing down and cooling near the wall.

3.3 Gas separation in the Knudsen pump

At the present time, there is an increasing interest in the micro devices driven by the effect of thermal transpiration and operating as pumps [17]. In such devices flow of gas mixture can be even more interesting [18], [19] since the magnitude of flows induced by the thermal transpiration depends on properties of the gas molecules. Individual mixture components may behave differently inside the micro devices.

Here we present results of gas separation modeling in the Knudsen pump. For the purpose of reducing of the computational time and increasing the accuracy of the results the problem has been considered in plane geometry. Examples of 3D modeling of gas flows in the Knudsen pump and its modifications carried out on our PSE can be found in [5] and [20] for a simple gas and a gas mixture, respectively.

The scheme of the Knudsen pump is shown in Fig. 8. The pump consists of two reservoirs filled with a gas, which are sequentially connected by a narrow and wide channels. The width of channels are h and H, respectively. At the reservoirs walls the temperature T_1 is kept, along the walls of the narrow channel the temperature increases from T_1 to T_2 at the joint of channels, and then decreases along the wall of the wide channel to the T_1. The condition of the diffuse molecule reflection is set at the reservoirs and channels walls. The construction of the pump is symmetric about the channel axis; this allows one to carry out calculations only in a half of the system with the specular boundary condition on the symmetry line.

A two component mixture of hard sphere molecules having masses m^α and m^β is considered. At each point of the device the initial gas temperature are equal to the temperature of the nearest walls. The initial concentrations of the components are equal, $n^\alpha = n^\beta$. The initial pressure of the mixture in the pump is constant.

Let us consider the processes of pumping and gas separation in the system with the following parameters: $m^\alpha/m^\beta = 1/2$, $T_2/T_1 = 4/3$, $H = 2h$, (Kn $= 0.5$, $\lambda = 1/\sqrt{2}\pi\sigma^2(n^\alpha + n^\beta)$. Fig. 9 shows time evolutions of the total pressure ratio and partial pressure ratios of the components. Under the action of the effect of thermal creep the gas starts flowing in the near-wall area along the temperature gradient. Then since the thermal creep effect is stronger in the narrow channel than in the wide one, the both gas components are pumped from the left to the right reservoir. At $t = 250\tau$ the ratio of the partial pressures of the light component reaches its maximal value, and the inverse flux of the light component driven by the increased total pressure in the right reservoir begins.

Fig. 10 shows the steady-state distribution of the gas pressure $p(x) = p^\alpha(x)+p^\beta(x)$ and relative concentrations of the light gas $\chi^\alpha = n^\alpha/(n^\alpha+n^\beta)$ and the heavy gas $\chi^\beta = n^\beta/(n^\alpha + n^\beta)$ along the axis of the pump. The pumping rate reaches the value $p_A/p_B \approx 1.015$, and the separation in the right reservoir is equal to $n^\alpha_B/n^\beta_B \approx 0.987$.

The gas flow in the pump is shown in Fig. 11. At the area near the channels walls the gas flows in the direction of the temperature gradient, at the middle of channels the reverse Poiseuille flow is formed due to the increased pressure at the joint.

3.4 Multistage Knudsen pump

The pumping and the gas separation level can be increased using the multistage Knudsen pump. In the multistage pump the reservoirs are connected by a cascade of sequentially connected narrow and wide channels. The temperature distribution at the channels is the same as in the single-stage pump: increases at the narrow channel and decreases at the wide one.

Consider the modeling of the multistage device for the Knudsen number Kn $= 0.5$. The temperatures ratio is equal to $T_2/T_1 = 2$, and $m^\alpha/m^\beta = 1/2$. In Fig. 11 the field of the relative concentration χ^α is shown. In the Fig. 12 the distributions of the total pressure and the relative concentrations of the components along the pump axis are given. The graphics have a saw-toothed form. The local maxima are located at the joints of the channels. The total pumping level is much larger compared to the single-stage device and equal to $p_A/p_B = 1.32$, and the total gas separation is equal to $\chi^\alpha_B/\chi^\alpha_A = 0.93$.

In addition, consider the gas mixture separation in the Knudsen pump using the Lennard-Jones potential of intermolecular interaction $U^{ij} = 4\varepsilon^{ij}((\sigma^{ij}/r)^{12} - (\sigma^{ij}/r)^6)$. Let the gas be a mixture of Neon and Argon for which parameters of the Lennard-Jones potential are: $\varepsilon^{Ne} = 35.7K$, $\varepsilon^{Ar} = 122.4K$, $\sigma^{Ne} = 2.789A$, $\sigma^{Ar} = 3.482A$. The interaction parameters for different atoms are defined by the combinatory relations: $\varepsilon^{\alpha,\beta} = \sqrt{\varepsilon^\alpha\varepsilon^\beta}$, $\sigma^{\alpha,\beta} = (\sigma^\alpha + \sigma^\beta)/2$. The Knudsen number is defined by the mean free path for Neon, $\lambda = 1/\sqrt{2}\pi(\sigma^{Ne})^2\Omega^{(2,2)}_{Ne}n_0$. An omega integral $\Omega^{(2,2)}_{Ne} = 0.8135$ is used to make equal the viscosity of the Lennard-Jones gas to the viscosity of the hard sphere gas.

Fig. 13 shows a comparison of the distributions of total pressure and the relative

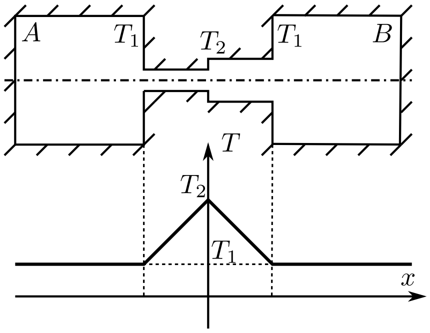


Figure 7: Scheme of the Knudsen pump.

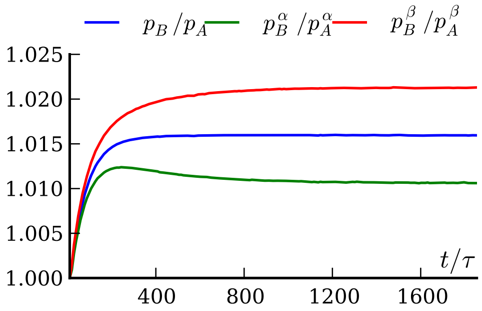


Figure 8: Pumping and gas separation versus time.

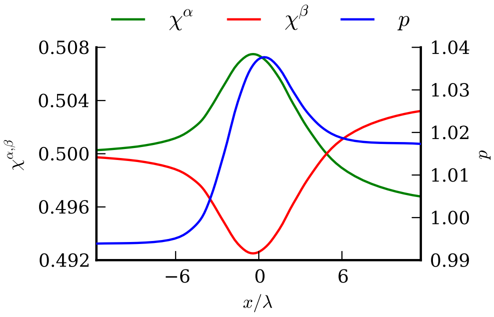


Figure 9: Pressure and relative concentrations along the pump axis.

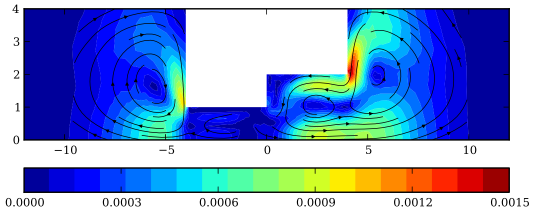


Figure 10: Streamlines of the gas flow

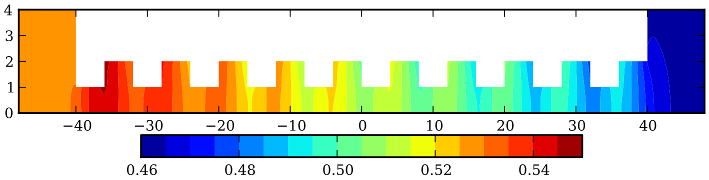


Figure 11: Field of the relative concentration χ^α

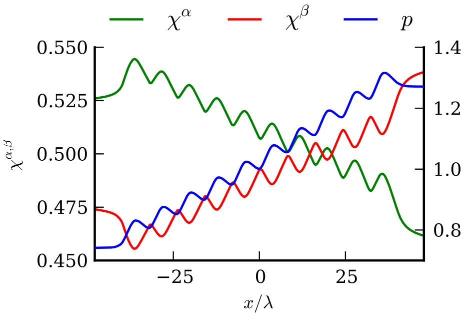


Figure 12: Pressure and relative concentrations along the multistage pump axis.

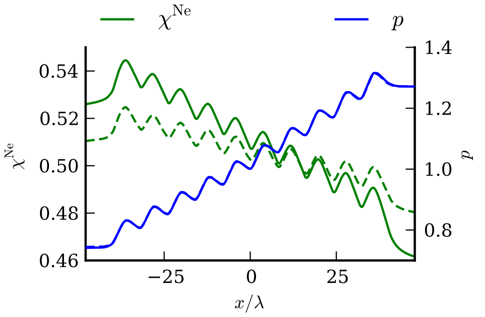


Figure 13: Pressure and relative concentration of light gas along the pump axis.

concentrations of the components along the pump axis. The solid line corresponds to the simulation using the hard-sphere potential, and the dashed line to the Lennard-Jones potential. The results for the pressure for the both considered potentials are very close, but the separation levels differ. For the Lennard-Jones potential the separation level is almost two times lower, however qualitatively in the both cases the processes evolve in the similar manner.

Conclusion

The description of the problem solving environment and the examples provided in this paper show that studies of gas flows in micro channels and micro devices can be efficiently done by using numerical modeling. The applied approach is based on the direct solution of the Boltzmann kinetic equation. Thus the accuracy of the results is limited only by the available computational resources. As the examples slightly disturbed flows as well as non-steady supersonic flows are presented. Flows of simple gas and gas mixture with realistic intermolecular potential are considered.

References

[1] Cheremisin, F. G. Physics – Doklady 1997; 42:607–610.

[2] Cheremisin, F. G. Doklady Physics 2000; 45(8):401–404.

[3] Tcheremissine, F. G. Comp. Math. and Math. Phys. 2006; 46(2):315–329.

[4] N. I. Khokhlov, Yu. Yu. Kloss, B. A. Shurygin, F. G. Tcheremissine. In: Proceedings of the 26th International Symposium on Rarefied Gas Dynamics. AIP Conf. Proc. 1084; 2008, p. 1039–1044.

[5] Anikin Yu.A., Derbakova E.P., Dodulad O.I., Kloss Yu.Yu., Martynov D.V., Rogozin O.A., Shuvalov P. V., Tcheremissine F.G. Procedia Computer Science 2010; 1(1):735–744.

[6] Yu. A. Anikin, O. I. Dodulad, Yu. Yu. Kloss, D. V. Martynov, P. V. Shuvalov, F. G. Tcheremissine. Vacuum 2012; 86(11):1770–1777.

[7] F. G. Tcheremissine. http://www.chemphys.edu.ru/pdf/2007-10-22-001.pdf, 2007 (in Russian).

[8] O. I. Dodulad, F. G. Tcheremissine. In: 28th International Symposium on Rarefied Gas Dynamics 2012, AIP Conf. Proc. 1501, p. 302–309.

[9] F. G. Tcheremissine. Comp. Math. and Math. Phys. 2012; 52(2), 252–268.

[10] Ivanov M.S. et al. In: Proc. of 25th Intern. Symp. on Rarefied Gas Dynamics. 2006, p. 21–28.

[11] V. I. Kolobov, R. R. Arslanbekov, V. V. Aristov, A. A. Frolova, S. A. Zabelok, F. G. Tcheremissine. Internal Symposium on Rarefied Gas Dynamics, Melville, N.Y; 2005, p. 96–101.

[12] Kloss Yu.Yu., Shuvalov P.V., Tcheremissine F.G. Procedia Computer Science 2010; 1(1): 1077–1085.

[13] *GMSH.* http://geuz.org/gmsh/.

[14] S. Varoutis, D. Valougeorgis, F. Sharipov. In: Proceedings of the 26th International Symposium on Rarefied Gas Dynamics. AIP Conf. Proc. 1084; 2008, p. 1111–1116.

[15] Titarev. Comm. in Comp. Phys. 2012; 52(2): 269–284.

[16] F. Seiler, B. Schmidt. In: Rarefied gas dynamics; International Symposium, 12th, Charlottesville, VA, 1981, p. 1094-1104.

[17] N. K. Gupta, Y. B. Gianchandani. Microporous and Mesoporous Materials 2011; 142:535–541.

[18] S. Takata, H. Sugimoto, S. Kosuge. European Journal of Mechanics - B/Fluids 2007; 26(2):155–181.

[19] H. Sugimoto, A. Shinotou. In: Proc. of 25th Intern. Symp. on Rarefied Gas Dynamics. AIP Conf. Proc 1333(1); 2011, p. 784–789.

[20] O. I. Dodulad, I. D. Ivanova, Yu. Yu. Kloss, P. V. Shuvalov, F. G. Tchremissine. In: Proc. of the 28th International Symposium on Rarefied Gas Dynamics. AIP Conf. Proc. 1501; 2012, p. 816–823.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Solving systems of nonlinear mixed Fredholm–Volterra integro–differential equations using fixed point techniques and biorthogonal systems

M. I. Berenguer[1], **D. Gámez**[1] and **A. J. López Linares**[1]

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

emails: `maribel@ugr.es`, `domingo@ugr.es`, `alopezl@ugr.es`

Abstract

Using fixed point techniques and biorthogonal systems in adequate Banach spaces, the problem of approximating the solution of a system of nonlinear mixed Fredholm–Volterra integro–differential equations of the second kind is turned into a new numerical method that allows it to be solved numerically.

Key words: Systems of nonlinear mixed Fredholm–Volterra integro–differential equations, biorthogonal systems, fixed point, numerical methods.
MSC 2000: AMS 45J05, 45L05, 45N05, 65R20.

1 Introduction

Some important problems in science and engineering can usually be reduced to a systems of differential, integral and integro-differential equations. Therefore, the importance of such systems in the study of problems arising from the real world, have made them an important research topic in Applied Mathematics. Since few of these equations can be solved analytically, it is often necessary to develop numerical techniques that allow us to obtain the approximate solution of such equations. The increasing success with which the numerical methods have been used to model concrete situations, have made them an important research topic in Numerical Analysis (see for example [1], [2], [3], [7], [9], [10], [11], [12], [14] and [15])

2 The problem

Let us consider the system of nonlinear Fredholm–Volterra integro–differential equations of the second kind:

$$\begin{cases} \mathbf{X}'(t) = \mathbf{F}(t, \mathbf{X}(t)) + \displaystyle\int_{\alpha}^{\alpha+\beta} \mathbf{K}(t, s, \mathbf{X}(s))ds + \int_{\alpha}^{t} \mathbf{H}(t, s, \mathbf{X}(s))ds, \qquad (t, s \in [\alpha, \alpha+\beta]) \\ \mathbf{X}(\alpha) = \rho. \end{cases} \tag{1}$$

where:

a) $\alpha \in \mathbb{R}$ and $\rho = [\rho_1, \ldots, \rho_n]^T \in \mathbb{R}^n$,

b) $\mathbf{X}(t) = [x_1(t), \ldots, x_n(t)]^T$ with $\mathbf{X} \in C([\alpha, \alpha+\beta], \mathbb{R}^n)$ is the solution to be calculated,

c) $\mathbf{F} \in C([\alpha, \alpha+\beta] \times \mathbb{R}^n, \mathbb{R}^n)$ with

$$\mathbf{F}(t, \mathbf{X}(t)) = [F_1(t, \mathbf{X}(t)), \ldots, F_n(t, \mathbf{X}(t))]^T,$$

d) $\mathbf{K}, \mathbf{H} \in C([\alpha, \alpha+\beta]^2 \times \mathbb{R}^n, \mathbb{R}^n)$ with

$$\mathbf{K}(t, s, \mathbf{X}(s)) = [K_1(t, s, \mathbf{X}(s)), \ldots, K_n(t, s, \mathbf{X}(s))]^T \quad \text{and}$$

$$\mathbf{H}(t, s, \mathbf{X}(s)) = [H_1(t, s, \mathbf{X}(s)), \ldots, H_n(t, s, \mathbf{X}(s))]^T.$$

In this work we present a numerical method to solve the above mentioned system based on two classical analytical tools: the fixed point theorem and biorthogonal systems in a Banach space. Such tools have been used also successfully in [4] and [5]. We recall some auxiliary facts from the Fixed Point Theory and the Theory of Schauder Bases, we present the method when the Lipschitz condition is assumed on **F**, **G** and **H** in their last variables and we obtain an explicit control of the error committed. Finally we illustrate the theoretical results with some examples.

References

[1] S. Abbasbandy and A. Taati, *Numerical solution of the system of nonlinear Volterra integro–differential equations with nonlinear differential part by the operational Tau method and error estimation*, J. Comput. Appl. Math. **231** (2009), pp. 106–113.

[2] A. Akyüz–Daşcioğlu and M. Sezer, *Chebyshev polynomial solutions of systems of higher–order linear Fredholm–Volterra integro–differential queations*, J. Franklin Inst. **342** (2005), pp. 688–701.

[3] A. Arikoglu and I. Ozkol, *Solution of integral and integro–differential equation systems by using differential transform method*, Comput. Math. Appl. **56** (2008), pp. 2411–2417.

[4] M. I. Berenguer, D. Gámez and A. J. López Linares, *Fixed–point iterative algorithm for the linear Fredholm–Volterra integro–differential equation*, J. Appl. Math. Vol. **2012** (2012), doi:10.1155/2012/370894. Article ID 370894, 12 pages.

[5] M.I. Berenguer, D. Gámez and A.J. López Linares, *Fixed point techniques and Schauder bases to approximate the solution of the first order nonlinear mixed Fredholm-Volterra integro-differential equation*, in press, J. Comput. Appl. Math. (2012), doi:10.1016/j.cam.2012.09.020.

[6] B. Gelbaum and J. Gil de Lamadrid, *Bases on tensor products of Banach spaces*, Pacific J. Math. **11** (1961), pp. 1281–1286.

[7] M. H. Heydari, M. R. Hooshmandasl, F. Mohammadi and C. Cattani, *Wavelets method for solving systems of nonlinear singular fractional Volterra integro–differential equations*, Commun. Nonlinear Sci. (2013), doi: http://dx.doi.org/101016/j.cnsns.2013.04.026.

[8] G. J. O. Jameson, *Topology and Normed Spaces*, Chapman & Hall, London, 1974.

[9] K. Maleknejad and M. Tavassoli Kajani, *Solving linear integro–differential equation system by Galerkin methods with hybrid functions*, Appl. Math. Comput. **159** (2004), pp. 603–612.

[10] K. Maleknejad, B. Basirat and E. Hashemizadeh, *A Berstein operational matrix approach for solving a system of high order linear Volterra–Fredholm integro–differential equations*, Math. Comput. Model. **55** (2012), pp. 1363–1372.

[11] J. Pour–Mahmoud, M. Y. Rahimi–Ardabili and S. Shahmorad, *Numerical solution of the system of Fredholm integro–differential equations by the Tau method*, Appl. Math. Comput. **168** (2005), pp. 465–478.

[12] J. Saberi–Nadjafi and M. Tamamgar, *The variational method: A highly promising method for solving the system of integro–differential equations*, Int. J. Comput. Math. **56** (2008), pp. 346–351.

[13] Z. Semadeni, *Product Schauder bases and approximation with nodes in spaces of continuous functions*, Bull. Acad. Polon. Sci. **11** (1963), pp. 387–391.

[14] E. Yusufoğlu, *An efficient algorithm for solving integro–differential equations system*, Appl. Math. Comput. **192** (2007), pp. 51–55.

[15] Ş. Yüzbaşı, N. Şahin and M. Sezer, *Numerical solution of systems of linear Fredholm integro–differential equations with Bessel polynomial bases*, Comput. Math. Appl. **61** (2011), pp. 3079–3096.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

A new use of correlation-immune Boolean functions in cryptography and new results

S. Bhasin[1], C. Carlet[2] and S. Guilley[1]

[1] *Institut Mines-Telecom, Telecom-ParisTech, CNRS LTCI (UMR 5141), 37/39 rue Dareau 75 014 Paris, France*

[2] *LAGA, Universities of Paris 8 and Paris 13; CNRS, UMR 7539, Department of Mathematics, 2 rue de la liberté, 93526 Saint-Denis cedex 02, France*

emails: `shivam.bhasin@TELECOM-ParisTech.fr`, `claude.carlet@univ-paris8.fr`, `sylvain.guilley@TELECOM-ParisTech.fr`

Abstract

We shall describe how correlation-immune functions and orthogonal arrays can play a new role in cryptography for optimizing the masking counter-measures to side-channel attacks. We shall show why the state of the art on correlation-immune functions does not allow addressing the new questions posed by this emerging framework. And we shall give some original results in this sense.

Key words: Correlation immune Boolean function, orthogonal array, side channel attack, masking

1 Extended abstract

Correlation-immune Boolean functions (from $\mathbb{F}_2^n$ to $\mathbb{F}_2$) and their related orthogonal arrays have well-known applications in symmetric cryptography, since they can be used in the pseudo-random generators of stream ciphers to combine the outputs to several LFSR (in the so-called combiner model); their correlation immunity allows resisting the Siegenthaler attack.

Definition A Boolean function f is dth-order correlation-immune (d-CI for short) if its output value distribution does not change when at most d coordinates of its input are fixed to arbitrary values.

The array whose rows are the elements of the support of f is an orthogonal array of strength d.

A nice characterization exists through the Walsh transform of the function.

Definition The Fourier transform $\widehat{f}$ of a Boolean function (viewed as integer-valued) or of any integer-valued function f over $\mathbb{F}_2^n$ is defined as $\widehat{f} : \begin{array}{l} \mathbb{F}_2^n \mapsto \mathbb{Z} \\ x \mapsto \sum_{x \in \mathbb{F}_2^n} f(x)(-1)^{a \cdot x} \end{array}$ where $a \cdot x$ is the usual inner product in $\mathbb{F}_2^n$. The Walsh transform W_f of a Boolean function f over $\mathbb{F}_2^n$ is the Fourier transform of its sign function $(-1)^f$.

It has been proved by Xiao and Massey that f is d-CI if and only if $W_f(a)$ is null for every nonzero vector a of weight at most d. See more in the survey on Boolean functions [1].

In the framework of stream ciphers, the functions must be balanced, that is, have uniform output distribution. Indeed, if such function is unbalanced, then the plaintext and the ciphertext output by the stream cipher have an obvious statistic dependency, which allows distinguishing when a pair of texts is a pair (plaintext, ciphertext). Correlation-immune balanced functions (called resilient functions) have been extensively studied until the invention in 2003 of algebraic attacks on stream ciphers. No construction of functions achieving resistance to the Siegenthaler attack and to algebraic and fast algebraic attacks is known yet. This makes the recent research in the domain of CI-functions has now more theoretical interest than practical.

We will show a new way of having correlation immune functions playing a role in cryptography; the functions in this framework will need to have low weight and then to be unbalanced. The implementation of cryptographic algorithms in devices like smart cards, FPGA or ASIC leaks information on the secret data, leading to *side channel attacks* (SCA). These attacks are very powerful if countermeasures are not included in the implementation of cryptosystems: recall that a cryptanalysis is considered breaking a crypto-system if it works in less that 2^{80} operations, needing thousands of centuries of computation with current computers and using thousands of pairs of plaintexts-ciphertexts, while SCA can be efficient, recovering the key with few plaintext-ciphertext pairs in a few seconds.

Counter-measures fortunately exist but are costly in terms of running time and of memory when they need to resist higher order side channel attacks. The most commonly used counter-measure is a secret-sharing method called *masking*. Some implementations of masking counter-measures can be very efficient but cannot afford using all the possible values of the mask variable. It can be shown that, for a given level of security, the selection of the mask values is optimal when it is chosen in the support of a d-CI Boolean function, where d is as large as possible. The lower the weight of this function, the cheaper the countermeasure.

This poses new questions on correlation-immune functions. The numerous studies made until now dealt with resilient functions and do not work for generating low weight correlation-immune functions. We will show why the usual constructions (such as the Maiorana-McFarland construction) and the usual ways of deriving new CI-functions from known ones (secondary constructions such as the indirect sum) are not efficient for generating low weight CI functions.

We use Satisfiability Modulo Theory (SMT) tools to study CI functions in at most 10 variables. This number is sufficiently large for cryptographic applications to block ciphers (more prone to SCA). We give the minimal weights of d-CI functions in n variables for $n \leq 10$ and all $d \leq n$. Some of these results were not known previously, such as the minimal weight for $(n = 9; d = 4)$ and $(n = 10; d \in \{4; 5; 6\})$. These results set new bounds for the minimal number of lines of binary orthogonal arrays.

We show that a byte-oriented block cipher such as AES can be protected with only 16 mask values against correlation power attacks of orders 1, 2 and 3.

References

[1] C. Carlet, "Boolean Functions for Cryptography and Error Correcting Codes", Chapter of the monography "Boolean Models and Methods in Mathematics, Computer Science, and Engineering," *Cambridge University Press* (Peter Hammer and Yves Crama editors), pages 257-397, 2010.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Product Type Weights Generated by a Single Nonproduct Type Weight Function in High Dimensional Model Representation (HDMR)

Derya Bodur[1] **and Metin Demiralp**[1]

[1] *Informatics Institute, Computational Science and Engineering Program, İstanbul Technical University*

emails: `deryabodur1@gmail.com`, `metin.demiralp@gmail.com`

Abstract

This work focuses on the utilization of the nonproduct type weight functions in HDMR. The problem of this type HDMR is the fact that the use of vanishing conditions on the components can not be imposed without causing inconsistencies. We have avoided these impositions in our previous studies in bivariate and trivariate HDMR. What we have obtained there has been the fact that the vanishing conditions can be imposed under not the nonproduct type weight function but under the product of its univariate integrals to construct appropriate HDMR. This work extends this result to the most general multivariance, the HDMR of n variable multivariate functions.

Key words: High Dimensional Model Representation, Weight Functions, Weight Function Integrals

1 Introduction

Recently we have focused on the impositions to get uniqueness in the construction of HDMR under orthogonal geometry but nonproduct type weight function. The nonproduct type weight function destroys the possibility of using vanishing conditions employed in HDMRs under product type weight functions [1]. This happens because of some inconsistencies in the equations to determine the HDMR components. This fact has urged us not to impose the vanishing conditions. The ordinary HDMR under product type weight functions uses the equations of each possible level of multivariance starting from the constancy [2–15]. These equations can be produced by using just the HDMR equation and its univariate, bivariate and higher variate integrals up to and including the highest multiple integrations where HDMR is integrated over all independent variables. This procedure produces some number equations which are populated same as the HDMR components. Even though the linearity of the equations in HDMR components facilitate the analysis the integrals over them appear due to the integrations to construct the equations.

We have started with the simplest nontrivial case, bivariate HDMR for better understanding the nature of the problem under consideration. After some efforts we have arrived the fact that the HDMR components can be uniquely determined by the vanishing conditions under not the nonproduct type weight function but the product of its univariate integrals. We have soon noticed that this very pleasent conclusion was not peculiar to the bivariate HDMR and we have extended our studies to the trivariate HDMR. What we have found was exactly same conclusion except the number of the factors and the independent variables was three. We have reported these cases and extended our studies to the HDMR cases where the number of independent variables is four and higher, and, the weight function is again nonproduct type. The conclusion were what we have expected. However, these conclusion were incomplete because they could be only the beginning stage of the mathematical induction based proof. We now completed the circle and by assuming the validity of the conclusion obtained rather small and fixed number of independent variables for a general number and could be able to show that the conclusion is conserved for the case where the number of independent variables is one more than the one in the validity assumed case. The proof is rather comprehensive for a proceeding because of the typographical reasons. However exactly same steps and the logic in the case of bivariate case. Hence we give details for bivariance. Grasping the issue for this case dissolves the induction stage difficulties in mathematical induction based proof.

2 Nonproduct Weighted HDMR for Bivariate Functions

High dimensional model representation of a given bivariate function can be given through the following equation

$$\begin{aligned} f\left(x_1,x_2\right) = f_0 + f_1\left(x_1\right) + f_2\left(x_2\right) + f_{12}\left(x_1,x_2\right), \\ x_1 \in \left[\, a_1, b_1 \,\right], \quad x_2 \in \left[\, a_2, b_2 \,\right] \end{aligned} \tag{1}$$

where f symbolizes the given bivariate function while the independent variables are denoted by x_1 and x_2. Because of the nature of the HDMR, the right hand side of (1) contains four components: a constant (f_0), two univariate functions (f_1 and f_2), and, a single bivariate function (f_{12}). These components can not be uniquely determined unless certain conditions are imposed on them. To understand how these impositions can be realized we can multiply both sides of (1) by a bivariate weight function denoted by $\Omega\left(x_1,x_2\right)$ (the weight function can only be at most finitely many times vanishing in the region defined as $\left[\, a_1, b_1 \,\right] \times \left[\, a_2, b_2 \,\right]$

by definition) and then separately integrate with respect to x_1 and x_2 over their domains. These give

$$\begin{aligned}\int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f\left(x_1,x_2\right) &= \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f_0 + \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f_1\left(x_1\right) \\ &+ \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f_2\left(x_2\right) \\ &+ \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f_{12}\left(x_1,x_2\right), \quad x_1 \in \left[\, a_1, b_1 \,\right]\end{aligned} \tag{2}$$

$$\begin{aligned}\int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right) f\left(x_1,x_2\right) &= \int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right) f_0 + \int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right) f_1\left(x_1\right) \\ &+ \int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right) f_2\left(x_2\right) \\ &+ \int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right) f_{12}\left(x_1,x_2\right), \quad x_2 \in \left[\, a_2, b_2 \,\right]\end{aligned} \tag{3}$$

For simplicity we assume that the following unit integral condition is satisfied by the weight function.

$$\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) = 1 \tag{4}$$

The very specific right hand side structures of (2) and (3) urge us to define

$$\Omega_1\left(x_1\right) \equiv \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right), \qquad \Omega_2\left(x_2\right) \equiv \int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right) \tag{5}$$

for which the following unit integral conditions remain valid.

$$\int_{a_1}^{b_1} dx_1\Omega_1\left(x_1\right) = 1, \qquad \int_{a_2}^{b_2} dx_2\Omega_2\left(x_2\right) = 1 \tag{6}$$

Apparently Ω_1 and Ω_2 are univariate weight functions which convert (2) and (3) to the following concise forms

$$\begin{aligned}\int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f\left(x_1,x_2\right) &= \Omega_1\left(x_1\right) f_0 + \Omega_1\left(x_1\right) f_1\left(x_1\right) + \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f_2\left(x_2\right) \\ &+ \int_{a_2}^{b_2} dx_2\Omega\left(x_1,x_2\right) f_{12}\left(x_1,x_2\right), \qquad x_1 \in \left[\, a_1, b_1 \,\right]\end{aligned} \tag{7}$$

$$\begin{aligned}\int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right)f\left(x_1,x_2\right) \quad &= \quad \Omega_1\left(x_1\right)f_0+\Omega_1\left(x_1\right)f_1\left(x_1\right)+\int_{a_1}^{b_1} dx_1\Omega\left(x_1,x_2\right)f_2\left(x_2\right)\\ &\qquad +\int_{a_1}^{b_1} dx_2\Omega\left(x_1,x_2\right)f_{12}\left(x_1,x_2\right), \qquad\qquad x_2\in\left[\,a_2,b_2\,\right]\end{aligned} \tag{8}$$

If we define

$$D\left(x_1,x_2\right)\equiv\Omega\left(x_1,x_2\right)-\Omega_1\left(x_1\right)\Omega_2\left(x_2\right) \tag{9}$$

then we can easily show that its univariate integrals over the independent variables' domains vanish independently. Hence it somehow measures the deviation between the given nonproduct type weight function and the product of its univariate integrals. By using $D\left(x_1,x_2\right)$ we can write the following equation

$$\Omega\left(x_1,x_2\right)=\Omega_1\left(x_1\right)\Omega_2\left(x_2\right)+D\left(x_1,x_2\right). \tag{10}$$

Now we can rewrite original HDMR equality in homogeneous form and then multiply the both sides by the product $\Omega_1\left(x_1\right)\Omega\left(x_2\right)$ and independently integrate with respect to x_2 and x_1 over their intervals. These give

$$\begin{aligned}\Omega_1\left(x_1\right)\Bigg[\int_{a_2}^{b_2} dx_2\Omega_2\left(x_2\right)f\left(x_1,x_2\right)-f_0-f_1\left(x_1\right)-\int_{a_2}^{b_2} dx_2\Omega_2\left(x_2\right)f_2\left(x_2\right)\\ -\int_{a_2}^{b_2} dx_2\Omega_2\left(x_2\right)f_{12}\left(x_1,x_2\right)\Bigg]=\int_{a_2}^{b_2} dx_2D\left(x_1,x_2\right)f_2\left(x_2\right)\\ +\int_{a_2}^{b_2} dx_2D\left(x_1,x_2\right)f_{12}\left(x_1,x_2\right)-\int_{a_2}^{b_2} dx_2D\left(x_1,x_2\right)f\left(x_1,x_2\right)\end{aligned} \tag{11}$$

$$\begin{aligned}\Bigg[\int_{a_1}^{b_1} dx_1\Omega_1\left(x_1\right)f\left(x_1,x_2\right)-f_0-f_1\left(x_1\right)-\int_{a_1}^{b_1} dx_1\Omega_1\left(x_1\right)f_2\left(x_2\right)\\ -\int_{a_1}^{b_1} dx_1\Omega_1\left(x_1\right)f_{12}\left(x_1,x_2\right)\Bigg]\Omega_2\left(x_2\right)=\int_{a_1}^{b_1} dx_1D\left(x_1,x_2\right)f_2\left(x_2\right)\\ +\int_{a_1}^{b_1} dx_1D\left(x_1,x_2\right)f_{12}\left(x_1,x_2\right)-\int_{a_1}^{b_1} dx_1D\left(x_1,x_2\right)f\left(x_1,x_2\right)\end{aligned} \tag{12}$$

Not only the double integral of the deviation function $D\left(x_1,x_2\right)$, but also its all univariate integrals vanish as can be shown easily. This can be used to show that the right hand sides of the last two equations vanish. If we multiply the both sides of the HDMR equation's homogeneous version by $D\left(x_1,x_2\right)$ and independently integrate with respect to

x_1 and x_2 over their intervals then we can obtain equations which are equivalent to the right hand sides of the equations in (11) and (12). The vanishing right hand sides mean vanishing left hand sides where one of the two factors is a weight function. Therefore the expressions between the brackets in the left hand side of (11) and (12) must also vanish. Thus we can arrive at the following equations

$$\begin{aligned}\int_{a_2}^{b_2} dx_2 \Omega_2\left(x_2\right) f\left(x_1, x_2\right) &= f_0 + f_1\left(x_1\right) + \int_{a_2}^{b_2} dx_2 \Omega_2\left(x_2\right) f_2\left(x_2\right) \\ &\quad + \int_{a_2}^{b_2} dx_2 \Omega_2\left(x_2\right) f_{12}\left(x_1, x_2\right) \qquad (13)\end{aligned}$$

$$\begin{aligned}\int_{a_1}^{b_1} dx_1 \Omega_1\left(x_1\right) f\left(x_1, x_2\right) &= f_0 + \int_{a_1}^{b_1} dx_1 \Omega_1\left(x_1\right) f_1\left(x_1\right) + f_2\left(x_2\right) \\ &\quad + \int_{a_1}^{b_1} dx_1 \Omega_1\left(x_1\right) f_{12}\left(x_1, x_2\right) \qquad (14)\end{aligned}$$

Thus we have now three equation: original HDMR equation and these two univariate equations. We need one more equation at the level of constancy to complete the circle. To this end, we can multiply the both sides of the homogeneous version of the original HDMR equation by $\Omega(x_1, x_2)$ and then integrate the result over the domains of x_1 and x_2. What we obtain can be expressed as follows after using some of abovementioned definitions and entities

$$\begin{aligned}\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \Omega\left(x_1, x_2\right) f\left(x_1, x_2\right) &= f_0 + \int_{a_1}^{b_1} dx_1 \Omega_1\left(x_1\right) f_1\left(x_1\right) \\ &\quad + \int_{a_2}^{b_2} dx_2 \Omega_2\left(x_2\right) f_2\left(x_2\right) \\ &\quad + \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \Omega\left(x_1, x_2\right) f_{12}\left(x_1, x_2\right) \end{aligned} \qquad (15)$$

which can be rewritten in terms of the abovementioned deviation function as follows

$$\begin{aligned}&\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \Omega_1\left(x_1\right) \Omega_2\left(x_2\right) f\left(x_1, x_2\right) - f_0 - \int_{a_1}^{b_1} dx_1 \Omega_1\left(x_1\right) f_1\left(x_1\right) \\ &\quad - \int_{a_2}^{b_2} dx_2 \Omega_2\left(x_2\right) f_2\left(x_2\right) - \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \Omega_1\left(x_1\right) \Omega_2\left(x_2\right) f_{12}\left(x_1, x_2\right) \\ &\quad = \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 D\left(x_1, x_2\right) f_{12}\left(x_1, x_2\right) - \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 D\left(x_1, x_2\right) f\left(x_1, x_2\right)\end{aligned} \qquad (16)$$

whose right hand side vanishes as can be shown without difficulty. Thus we arrive at

$$\begin{aligned}\int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \Omega_1\left(x_1\right)\Omega_2\left(x_2\right) f\left(x_1,x_2\right) &= f_0 + \int_{a_1}^{b_1} dx_1 \Omega_1\left(x_1\right) f_1\left(x_1\right) \\ &+ \int_{a_2}^{b_2} dx_2 \Omega_2\left(x_2\right) f_2\left(x_2\right) \\ &+ \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \Omega_1\left(x_1\right)\Omega_2\left(x_2\right) f_{12}\left(x_1,x_2\right)\end{aligned} \tag{17}$$

The equations given in (1), (13), (14), and, (17) exactly match the equations obtained for the HDMR of $f\left(x_1,x_2\right)$ under the product type weight $\Omega_1\left(x_1\right)\Omega_2\left(x_2\right)$ over the rectangular geometry $\left[a_1,b_1\right]\times\left[a_2,b_2\right]$. Since the vanishing conditions are imposed for this specific HDMR we can also impose the following equations

$$\int_{a_i}^{b_i} dx_1 \Omega_i\left(x_i\right)\left[f_i\left(x_i\right) \;\; f_{12}\left(x_1,x_2\right)\right] = \left[0 \;\, 0\right], \qquad i=1,2. \tag{18}$$

This result can be expressed in the following statement: *One practically important way to use nonproduct type weight function in bivariate HDMR is to use the product of the univariate integrals of the weight function as the weight.* This can be considered as a theorem for the unitilization of nonproduct type weight functions in bivariate HDMR.

3 Generalization to Higher Dimensional Cases with Concluding Remarks

We have repeated the analysis of previous section for the trivariate functions and arrived at the same result. But this time a general trivariate weight function has been considered and we have reveled that the product of its univariate integrals can used as the product type weight function under which vanishing conditions can be imposed. This has been reported. In spite of not reporting we have found that the results obtained for bivariate and trivariate HDMR remain valid for the four, five, six and more variable HDMRs. However, it is not the way to show the validity for each number of variables. These findings can be used the starting stage of a mathematical induction proof. So we can consider the case where the number of the independent variables is n and assume the validity of the statement: *The HDMR of n variable multivariate functions under a hyperprismatic geometry but nonproduct type weight function can be set equivalent to the HDMR of same functions under the same geometry but under the product type geometry whose univariate weight factors are the univariate integrals of the nonproduct type weight function under consideration*, and then, try to prove the validity of same thing for $n+1$ variate HDMR in the same category. The proof is rather comprehensive in the formulation even though the intermediate steps are not hard to be taken. What we can say is that the proof can be obtained by following certain deviation functions based on the univariate and multivariate integrals of the nonproduct type weight function. We do not intend to get into details here for typographical reasons. However we are going to report the proof in a coming publication as soon as possible.

References

[1] Derya Bodur, Metin Demiralp, *A Perturbative Approach to High Dimensional Model Representation (HDMR) Under Nonproduct Type Weight and Over Orthogonal Geometry*, AIP Proceedings for the 10th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM2012), Kos, Greece (2012) 1994–1997.

[2] Sobol, *Sensitivity Estimates for Nonlinear Mathematical Models*, English Translation: MMCE **1** (1993) 407–414.

[3] Alış, Ö. F.; Rabitz, H., *General Foundations of High Dimensional Model Representation*, Journal of Mathematical Chemistry **25** (1999) 197–233.

[4] Alış, Ö. F.; Rabitz, H., *Efficient Implementation of HDMR*, Journal of Mathematical Chemistry **29** (2001) 127–142.

[5] Li, G.; Rosenthal, C.; Rabitz, H., *High Dimensional Model Representation*, Journal of Physical Chemistry **105** (2001) 7765–7777.

[6] Li, G.; Schoendorf, J.; Ho, T.; RabitZ, H., *Multicut-HDMR with an Application to an Ionospheric Model*, Journal of Computational Chemistry **25** (2004) 1149–1156.

[7] Tunga, M. A.; Demiralp, M., *A Factorized High Dimensional Model Representation on the Nodes of a Finite Hyperprimatic Regular Grid*, Applied Mathematics and Computation **164** (2005) 865–883.

[8] Demiralp, M.; Demiralp, E., *Reductive Multilinear Array Decomopsition Based Support Functions in Enhanced Multivariance Product Representation (EMPR)*, The International Conference on Applied Computer Science-2010 (WSEAS), Malta (2010) 448–454.

[9] Yaman, İ.; Demiralp, M., *HDMR Approximation of an Evolution Operator with a First Order Partial Differential Operator Argument*, App. Num. Anal. and Comp. Math., Wiley CHV (2003) 287–296.

[10] Tunga, B.; Demiarlp, M., *A Novel Hybrid High Dimensional Model Representation (HHDMR) Based on Combination of Plain and Logarithmic High Dimensional Model Representations*, WSEAS 12-th International Conference on Applies Mathematics for Science and Engineering **1** (2007) 157–161.

[11] Demiralp, M., *A Gentle Introduction to High Dimensional Model Representation Under Nonproduct Type Weights*, in Proceedings of the 2nd International Conference on Applied Informatics and Computing Theory, September, 2011 (WSEAS) (2011) 113–118.

[12] Li, G.; Artamonov, M.; Rabitz, H.; Wang, S.-W.; Georgopoulos, P. G.; Demiralp, M., *High Dimensional Model Representations generated from low order terms lp-rs-hdmr*, Journal of Computational Chemistry **24** (2003) 647–656.

[13] Demiralp, M., *Importance and measurement of univariance in high dimensional model representation for multivariate analysis*, WSEAS Transactions on Computers **5** (2006) 1738–1744.

[14] Tunga, B.; Demiralp, M., *Constancy maximization based weight optimization in high dimensional model representation*, Numer. Algor. **52** (2009) 435–459.

[15] Okan, A.; Baykara, N. A.; Demiralp, M., *Weight optimization in enhanced multivariance product representation (empr)*, In AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2010) **1281** (2010) 1935–1938.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Computational Modelling of Meteorological Variables on Multicores and Multi-GPU Systems

Murilo Boratto[1], Pedro Alonso[1] and Domingo Gimenéz[2]

[1] *Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, Spain*

[2] *Departamento de Sistemas Informáticos, Universidad de Murcia, Murcia, Spain*

emails: `muriloboratto@uneb.br`, `palonso@dsic.upv.es`, `domingo@um.es`

Abstract

The National Institute of Meteorology (INMET) in Brazil includes a network of meteorological stations which has 24 stations operating in agricultural region of São Francisco River Valley. There are 11 conventional stations and 13 with automatic systems. The institute gathers high quality expensive data despite they have been little used. Since there is no possibility of installing meteorological stations everywhere, a cross-validation procedure has been developed in order to evaluate the meteorological data from INMET network using inverse power distance weighting method. This work aims to present methodologies for meteorological variables representation through High Performance Computing Models on Multicore and Multi-GPU systems.

Key words: Computational Modelling, Meteorological Variables, Parallel Computing, Performance Estimation, Multicore, Multi-GPUs.

1 Introduction

Agricultural region of São Francisco River Valley has a successful public policy for the development of semiarid northeast region based on irrigated agriculture. Agriculture has stimulated this region of the country, bringing socio-economic development to the countryside in a wide area of modern agriculture. This development center is located in the driest area in the Northeast of Brazil, on the São Francisco River Valley, including a wide area, where natural conditions are exceptional for the development of irrigated fruit growing.

Efficiency is demanded in the usage of water for irrigated agriculture so that the process remains sustainable in environmental aspects, and the water resources management is the way one seeks to consider and resolve issues of scarcity. Meteorological variables representation provides data to quantify the water needs of crops, which is essential for its proper management. The main meteorological variables are: temperature, air relative humidity, wind velocity, solar radiation, precipitation and atmospheric pressure. The values of those variables are valid for the points where they are measured [1]. The National Institute of Meteorology (INMET) [2] in Brazil includes a network of meteorological stations which has 24 stations operating in São Francisco River Valley agricultural region. There are 11 conventional stations and 13 with automatic systems. These data are available on the Internet and have been little used. They are high quality data whose acquisition is considered expensive. Since there is no possibility to install meteorological stations everywhere, this project aims to test and validate, using the Inverse Power Distance Weighting [3], a methodology for making a meteorological data representation in INMET stations network.

The problem of this case study is the high computing power required to estimate meteorological variables that represent the climate of the São Francisco River Valley. The spatial interpolation method [4] requires a significant computational power to perform the interpolation on a large set of data, and sequential programming is not practicable in some cases. Currently it is usual to have computational systems formed by a multiprocessors with one or more Graphics Processing Units (GPUs). These systems are heterogeneous, due to different types of memory and different speeds of computation between CPU and GPU cores. In order to accelerate the solution of complex problems it is necessary to use the aggregated computational power of the two subsystems [5]. Heterogeneity introduces new challenges to algorithm design and system software. Our approach is to fully exploit all the CPU and all GPUs devices on these systems [6] and achieves the objective of maximum efficiency by an appropriate balancing of the workload among all the computational resources. Those ideas introduced methodologies for meteorological variables representation using the Inverse Power Distance Weighting method on Multicore and Multi-GPU systems.

2 Experimental Results

The computer used in our experiments has two 3.33 GHz Intel Xeon X5680 and 96GB GDDR3 main memory. Each one is an hexacore processor with 12 MB of cache memory. It contains two GPUs NVIDIA Tesla C2070 with 14 stream multiprocessors (SM) and 32 stream processors (SP) each (448 cores in total). In order to validate the presented methodology and derived equations, it will be applied computing techiques for efficient solution using inverse power distance weighting method by using the parallel scheme to represent the computational modeling of meteorological variables of São Francisco river valley region. We have implemented a parallel algorithm using OpenMP + CUDA. The benchmark was compiled with *nvcc*. In the experiments, we increased the number of stations from 12 to 18 to obtain the number that minimizes time. The execution with the original code is denoted by “Sequential”. As can be seen, its behavior is presented in Table 1.

Table 1: Comparative performance analysis of execution time of Sequential (in seconds).

Number of Estations	Automatic	Loop	Process
12	988.12	1,996.19	2,984.31
14	4,472.69	21,243.02	25,715.71
16	12,004.17	63,424.38	75,428.55
18	81,498.14	122,191.27	203,689.40
12	33.11%	66.89%	100.00%
14	17.39%	82.61%	100.00%
16	15.91%	84.08%	100.00%
18	40.01%	59.99%	100.00%

Where `Automatic Time` is the time for picks up the access time to the database, `Loop Time` refers to the internal loop (*while (p))*, which is the GPU has parallelized, and the `Process Time` includes the time to access the `Automatic + Loop`. The versions denoted by “Hybrid” represent executions denotes the use of several CPU cores and in two devices simultaneously.

Table 2: Comparative performance analysis of OpenMP + CUDA execution time (in seconds).

Number of Estations	Automatic	Loop	Process
12	284.83	157.81	442.58
14	912.75	1,250.66	2,154.24
16	4,283.81	2,458.08	6,742.75
18	14,498.81	7,723.21	22,222.84
12	61.23%	34.31%	100.00%
14	58.50%	41.50%	100.00%
16	66.72%	33.02%	100.00%
18	65.36%	34.64%	100.00%

Table 3: Comparative performance analysis of Speedup Rate.

Number of Estations	Automatic	Loop	Process
12	3.48	12.65	6.49
14	4.90	16.98	11.82
16	5.26	15.22	11.17
18	6.62	15.82	9.16

In this model, processes are executed by all the elements of the machine, the suitable number of CPU cores and the two GPU. The results of the experiments show that the parallel CPU + GPU algorithm reduces the execution time significantly. As can be seen in Table 3 the maximum speedup is around 6 for `Automatic Time` and 15 for `Loop Time`, matching the number of cores.

3 Conclusions and Future Work

From the results obtained, installation of new automatic meteorological stations or redistribution of already installed in São Francisco river valley is recommended in order to encourage the use of space interpolation, given its importance for Brazilian agriculture scenario. Furthermore, the performance results obtained in this work indicate that the proposed applications are effective, based on the classification task, response time and workload distribution. The applications built also demonstrate the proper measure of computing power that must be combined with intrinsic parallelism, which should be used in algorithm execution. Nevertheless, there are other models of distributed programming that can exploit the subject discussed in this article making it more complete. In this sense, new experiments have been developed to consolidate the mathematical model proposed. Nevertheless, it is still interesting to see the extension of such tuning technique to heterogeneous clusters for this particular problem.

References

[1] R. G. Allen, L. S. Pereira, D. Raes, M. Smith, Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage, Irrigation and Drainage 300 (56) (1998) 1–15.

[2] F. Oliveira, INMET: 100 Years of Weather in Brazil, INMET, 2009.
URL `http://books.google.com.br/books?id=VXfxZwEACAAJ`

[3] G. Y. Lu, D. W. Wong, An adaptive inverse-distance weighting spatial interpolation technique, Computer Geosciences 34 (9) (2008) 1044–1055.

[4] A. A. E. Jakob, A. F. Young, The use of interpolation methods in the analysis of spatial data sociodemographic, Brazilian Journal Sociodemographic.

[5] M. Boratto, P. Alonso, C. Ramiro, M. Barreto, Heterogeneous computational model for landform attributes representation on multicore and multi-GPU systems, Procedia CS 9 (2012) 47–56.

[6] F. Song, S. Tomov, J. Dongarra, Efficient support for matrix computations on heterogeneous multicore and multi-GPU architectures., Tech. Rep. 250, LAPACK Working Note (Jun. 2011).

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Impact of the Information Exchange Policies in Load Balancing Algorithms

Jose Luis Bosque[1], Oscar D. Robles[2], Pablo Toharia[2] and Luis Pastor[2]

[1] *Department of Electrónica y Computadores, Universidad de Cantabria*

[2] *Department of ATC and CCIA, Universidad Rey Juan Carlos*

emails: joseluis.bosque@unican.es, oscardavid.robles@urjc.es, pablo.toharia@urjc.es, luis.pastor@urjc.es

Abstract

This paper presents a detailed study about the information exchange rule of load balancing algorithm. It is focused on analyzing how this rule affects the performance as well as the scalability of these type of algorithms. To do that, different rules widely used in the bibliography have been implemented, as global rules but also as local rules. To compare them, a metric that looks for a compromise between the communications overhead and the speedup of the load balancing algorithm has been used: the Information Efficiency.

Key words: Heterogeneous computing, load balancing, load index.

1 Introduction

One of the most important challenges that supercomputers which are in the exascale computing route face is scalability. Nowadays, supercomputers have hundreds of thousands of cores, with a tendency for this quantity to grow in the future [1]. Analyzing and solving all the problems that affect the scalability of these systems will be one of the keys to make them run real applications with a high degree of efficiency.

One of the key aspects that have a deeper impact on the performance of the applications running on these supercomputers is the ability to achieve a good load balancing among all the available nodes. Load balancing algorithms try to assign each node a workload proportional to its computing capabilities [9]. In order to decide when a load balancing operation should be done these algorithms need accurate and updated information about the state of the system's nodes. The procedure to collect and update this information, called *Information Exchange Rule* [10], can introduce a great overhead in the system, specially when the number of available nodes grows. The scalability of this type of algorithms is greatly affected by this fact.

In order to completely define the information exchange rule two parameters are needed: a temporal distribution, i. e. when the operations of exchange of information take place, and a spatial distribution, i. e. which nodes are involved in the exchange of information. Regarding the temporal distribution, three types of rules are typically used:

- *On-demand.* The exchange of information takes place only when a node decides to perform a load balancing operation. At that moment state information from the rest of the nodes is requested. This way, the state information is always updated and the number of messages in the network is minimized, although the overhead associated to a load balancing operation is increased [8].

- *Periodical.* Each node periodically reports its state to the others, regardless of whether it is necessary or not. The selection of the best reporting frequency is critical [6].

- *Event-driven.* A node only reports its state when an event takes place. In this context, an event will be a node changing its state. This approach reduces the amount of communications and minimizes the risk of network saturation [5].

From an spatial point of view, these information rules can be *global* rules or *local* rules. Global rules define information exchanges with all the available nodes in the system. The information they manage is more complete, but they are not quite scalable [7]. Local rules are based on the concept of domain. A domain is a set of nodes among which the information is exchanged [2]. Domains can be isolated (one node can not appear in more than one domain) or overlapped (the domains share some of the nodes).

This paper studies how the information exchange rule affects the scalability of the load balancing algorithms. For that purpose, a basic load balancing algorithm has been implemented [4] and different information exchange approaches have been tested: periodical, event-driven and on-demand. Also, separated domains as well as overlapped domains have been implemented, and the results achieved in all cases are presented.

In order to do a fair comparison of all the affected parameters, a metric called information efficiency, proposed in [3] has been used. This metric takes into account a trade-off between the use of the network and the workload balancing achieved in the system. This trade-off is reflected in a better use of the available resources and therefore in a higher speedup achieved when the algorithm is used.

This paper is organized as follows. Section 2 presents a the proposed algorithm describing in detail the different information rules implemented. The experimental results are discussed in section 3. Finally section 4 exposes the conclusions and future work.

2 Design of the Load Balancing Algorithm

The approach presented in this paper is a dynamic, distributed, and non preemptive load balancing algorithm. It is a dynamic algorithm because the assignment of a task to an specific node is performed in run time. Then the task is completely executed in the assigned node, without any kind of task migration. It is a distributed algorithm since every node in the cluster takes its own decisions based on local stored information. The algorithm does not present any overhead if the nodes are naturally balanced, therefore it can automatically turn itself off on global under-loading or over-loading situations.

Depending on whether or not the algorithm uses domains, the decision about doing a load balancing operation can be global or local. When there are not any domains, once a node dedices to perform a load balancing operation the partner is selected among all the available nodes. On the other hand, the partner selection will be restricted to the nodes in the same domain when these are implemented.

2.1 Measuring the state of a node

During this phase the local information needed to determine the workload of the local node is collected, to calculate then the load index, which determines the state of the node. The decision about the local execution of a new task or launching a new load balancing operation is taken based on this state for each node. The load index is periodically computed and it is discretized in a set of states [4].

Load Index. The load index is calculated based on two static parameters, the number of cores and the computational power as well as a dynamic parameter, the number of tasks being currently executed in the node. A node has two sources of heterogeneity, the number of cores and their computational power. Thus, it is necessary to specify two different possibilities:

- The number of tasks is lower than the number of cores in the node. Therefore, there are some free cores and this node can accept more tasks, so it will be a *recipient* node.
- The number of tasks is larger than the number of cores. In this case, the load index is calculated based on the computational power of the nodes.

States of a Node. The states come from a discretization of the load index in order to minimize the exchange of global information, as well as to simplify load balancing decisions. These states determine the behaviour of a node, and three different ones have been defined:

- *Recipient State*: The state of a node is *recipient* when the number of its running tasks is less than the number of cores it has, or when its load index value is bigger than the threshold *neutral-recipient*. This means that the node has free cores and then it can assume at least one more task, either local or remote.

- *Neutral State*: Its load index has a medium value. In this case all the node's cores are executing at least one task. In this state the node can assume new local tasks but it rejects all remote requests.

- *Emitter State*: A node is Emitter when its load index has a value under the Neutral-Emitter threshold. This means that the node has many more tasks than the number of cores it has, and so it can not accept any more tasks.

Change of state. A node will change its state whenever a new measurement of the load index crosses a status threshold. Since this parameter is very volatile, it is possible that a particular node might be continuously changing its state due to small changes in its external load. To prevent this problem, some degree of hysteresis has been added to each threshold. This way, the number of messages due to state changes is reduced.

2.2 Global Information

In order to take decisions about load balancing it is necessary to exchange the state information among the nodes of the cluster. Our approach is a global algorithm, so all the nodes keep updated information about the global system state. Three different policies have been implemented for the nodes to broadcast their workload information when they suffer a change of state: periodical, event-driven and on-demand:

Periodical. The current state of a node, as well as the previous state of the rest of the nodes is sent after a certain period of time. Fixing the most appropriate period of time is of great importance, since a too short interval would mean sending messages so as to saturate the network but a too long interval would mean that the information that a node has about the others is obsolete. In order to avoid the effect of more than one change of state of one node, when a node receives information about another one it checks if the node the information is about is already in the recipients list, so the same node is not added or removed from the list more than once. With this implementation the new state of a node, as well as its old state, is received. The node will only be added to the list if the node changes from *recipient* to *neutral* state, and will only be removed from the list if it changes from *neutral* to *receptor* state, not being needed any change in the list in the rest of situations.

Event-driven. The information is sent to the rest of the nodes only when an event takes place, being an event a change of state on one node. The information sent is the current state of the node, since every change of state is communicated. As it was already mentioned, the algorithm is *emitter-initiated* so it is only of interest to know which are the *recipient* nodes. Therefore, an event will take place only when the node becomes *recipient* or when it leaves the *recipient* state. On the reception of a message showing a change of state, each node updates its receptors list. If the change is about a node that becomes *recipient*, that node is added to the list. Otherwise, it will be about a node that leaves the *recipient* state, so it will be removed from the list.

On-demand. State information will only be exchanged when a node is going to initiate a load balancing operation. In that moment the rest of the nodes are requested about their current state in order to have an updated list of possible *recipients*. If the information received from one node shows it is in *recipient* state, that node is added to the *recipients* list. It can be seen that every time a load balancing operation is going to be initiated, a new *recipients* list is created and any other previous list is deprecated, so the information is completely updated.

As it has been mentioned, load balancing operations can only take place between *Recipient* and *Emitter* nodes. Hence, only changes to or from *Recipient* state are significant enough to be communicated and thus, the number of messages is significantly reduced. Each node maintains a state-queue with the information received from other nodes. Only a *Recipient-queue* is needed, because only *Recipient* nodes can accept remote tasks. When a node becomes a *Recipient*, it broadcasts a message to all the nodes of the cluster so each of them will place it at the end of its queue. On the other hand, when a *Recipient* node changes its state to *Neutral* or *Emitter*, it broadcasts a message too and all the cluster nodes will discard it from their state-queues.

2.3 Local Information

When the number of nodes in the system grows, a global information scheme can be a problem from the point of view of the system's scalability. To avoid it, the nodes can be organized in domains, so only the nodes belonging to the same domain will exchange state information. Two different type of domains have been implemented: overlapped and random.

Overlapped domains. The nodes are assigned to a domain based on their identifier. There is an overlapping, so every node will be assigned to more than one domain. The overlapping is designed so as to be half the size of the domain, so each node will belong to 2 domains. The overlapping avoids the coexistance of overloaded and underused domains, where a great part of the workload is enclosed into a domain while there are others unoccupied. In this way, the common nodes between two different domains can act as bridges to move a certain part of the workload.

Random domains. The nodes are randomly organized in domains, son it is not needed to create any list as it happens in the previous case. Any time some information is going to be sent to a domain, the nodes that are going to receive it are randomly chosen. As time passes by all the nodes in the system will have enough information to perform their load balancing operations if needed.

2.4 Initiation of Load Balancing Operations

It determines when a new load balancing operation has to begin. Since a load balancing approach that does not interrupt any execution has been proposed, the decision about which node is finally going to execute a task can only be taken in the exact moment of beginning the execution. Therefore, there must be a *emitter-iniciated* rule to take this decision, and the decision about initiating a load balancing operation is *completely local* to the emitter. It must be noticed that on the decision of not doing any load balancing operation, no messages have to be exchanged, so the communications overhead is reduced. The initiation rule must be evaluated every time a new task has to be launched. At this moment, the state of the node is checked, in order to know if it can accept the local execution of the new task or if the search of a better candidate for the execution of the task should be initiated. A node can only accept the local execution of a new task if it is in *recipient* or *neutral* state. Therefore, only if the state of the node is *emitter*, a load balancing operation is initiated.

2.5 Partner Localization and Load Distribution

Once the decision of doing a load balancing operation is taken, two important steps must be performed: to locate a partner to send it the exceeding load and to decide how many tasks should be sent to that partner.

Partner localization. This is a completely local operation, since an *emitter* node looks for a partner in its own *recipients* queue. The selection of the partner can be done in many different ways. The approach adopted is to do a first random selection of a few nodes, that are then sorted based on their load index, so the less loaded node is requested. If the request is rejected, the next node in the list will be requested and so on. This strategy has some advantages. It reduces the communications overhead since the load indexes do not have to be constantly updated. Also, it avoids that one candidate is simultaneously chosen by different nodes, because it is in a prominent position in the queue.

Load Distribution. The last step to perform the load balancing operation is to decide how much workload should be sent to the *recipient* node. A good distribution is that in which each node gets an amount of workload that is proportional to its computational power. Therefore, the more similar load indexes are after the load balancing operation, the better the load distribution is. In this approach, first of all the *recipients* are sorted based on their load indexes. Then the *emitter* node sends to the first node in the sorted list as much workload as needed to force its load index to change from *recipient* to *neutral* state. Then, the second *recipient* from the list is selected and the same operation is performed. This process finishes when there is no more workload to send or there are no more *recipients.*

3 Experimental Results

The experiments have been run on a cluster composed of 8 nodes AMD single core. A total of 160 tasks to be delivered are sent to 2 nodes (80 tasks each), in order to test the algorithm but not to face an excessively large running time. This two nodes begin to execute tasks and as soon as their state changes to *emitter* they will send tasks to the rest of the available nodes in the system.

In order to test both global and local information algorithms, the following measures have been taken:

- t_{OP} is the time to execute the 160 tasks without any load balancing algorithm sending 20 of the 160 tasks to each node in the system.
- t_{DEL} is the time passed since the beginning or the test until all the tasks have been delivered (notice that it is not needed that the execution of the tasks have been completed).
- t_{EXE} is the time passed since the execution of the test begins until the execution of all the tasks launched is finished.
- n_{SEND} is the count of all the comunnications done among the nodes (global process) to inform about their state.
- lb_{CO} is the number of load balancing operations requested that are finally committed.
- lb_{NCO} is the number of load balancing operations that were requested but never were committed.
- ε is the information efficiency, measured as $\varepsilon = \frac{lb_{CO}}{m_{SEND} \times num\,of\,tasks} \times \frac{t_{OP}}{t_{EXE}}$

To carry out the experiments two types of static domains have been conformed: the first one with 4 nodes per domain with an overlapping of 2 nodes; the second one with 2 nodes per domain with an overlapping of 1 node. To carry out the experiments with random domains, 4 setups have been tested from 2 nodes per domain to 7 nodes per domain. Different tables with all the results achieved have been included, as well as two graphs collecting results from all the experiments.

Table 1 shows results from the experiments carried out without any domains. It means that the node 0 can deliver tasks to any of the available nodes. The other following tables collect results achieved using the different metrics explained in this paper. From these data figures 1 and 2 have been produced, and will be analyzed in the following. It must be said that this paper presents preliminary results, since the number of nodes used in the experiments is not big enough extract definitive conclusions about the advantages of using domains instead of global policies.

Table 1: Experimental results with global policies.

	t_{OP}	n_{SEND}	lb_{CO}	lb_{NCO}	t_{DEL}	t_{EXE}	t_{op}/t_{EXE}	ε
Event-driven	396	1960	121	161	395	432	0.917	3.54E-4
Periodical 1	396	22981	123	208	414	452	0.876	2.93E-05
Periodical 3	396	4648	124	476	419	457	0.867	1.45E-4
Periodical 5	396	4704	124	532	424	462	0.857	1.41E-4
Periodical 10	396	2576	118	1077	463	501	0.79	2.26E-4
On-demand	396	5544	123	500	393	430	0.921	1.28E-4

Global policies It can be seen in Figure 1 that the ratios t_{op}/t_{EXE} are quite similar for the on-demand, event-driven and periodical (with 3 seconds of updating time) policies. However, attending to the information efficiency the event-driven policy has a great advantage with respect to the other ones (see Figure 2). It can be seen that the execution times achieved applying this policy are very similar but the number of messages exchanged is much more reduced. This way, table 1 shows that applying this policy means the exchange of 1960 messages, in front of 4648 messages using the periodical policy and 5544 of the on-demand one. Since the number of load balancing operations performed is quite similar in all cases, it can be said that the event-driven policy achieves the same results but exchanging less information and therefore reducing the use of the network.

Another important aspect to consider when the different policies are compared is the number of load balancing operations that are rejected. The event-driven policy achieves 161 rejected operations, while the periodical policy shows 476 and the on-demand one 500. It also means a lower overhead so it can be said that applying the event-driven policy gives a better result.

Table 2: Experimental results with domains of 4 nodes.

	t_{OP}	n_{SEND}	lb_{CO}	lb_{NCO}	t_{DEL}	t_{EXE}	t_{op}/t_{EXE}	ε
Event-driven	396	1400	121	97	428	465	0.852	4.6E-4
Periodical 1	396	17080	121	99	432	469	0.844	3.74E-05
Periodical 3	396	5800	119	356	440	477	0.83	1.07E-4
Periodical 5	396	3640	120	436	459	497	0.797	1.64E-4
Periodical 10	396	1880	116	941	475	513	0.772	2.98E-4
On-demand	396	4225	121	103	428	465	0.852	1.52E-4

Finally, it must be noticed the deep impact that the election of the updating interval has on the periodical policy, since it can be seen that it produces a degradation of up to a 10%. This parameter is strongly dependent on the system's dynamism. In this experiment the tasks have an arriving rate on the nodes of 1 second. With it, policies with large updating interval are using obsolete information many times, something that means a high number of rejected load balancing operations (reaching 1077 in the case of periodical-10). However, in a less dynamic environment the policies more affected would be the ones that were using short updating intervals, since they would cause the exchange of a big amount of information that would never be used.

Static Domains With respect to the response time shown in Figure 1, once again the event-driven and on-demand policies achieve the best results, although the periodical-1 policy gives quite similar results. However, Figure 2 shows that attending to the efficiency, it is the event-driven policy the one that achieves much better results than the other ones. Basically the reasons are the same ones already explained in the previous section: it significantly reduces the number of messages exchanged and therefore the use of the network; provides a lower number of load balancing operations rejected, although in this case the difference with the other policies is lower, and its response times are very similar to the ones provided by the other policies.

Table 3: Experimental results with random domains of 3 Nodes.

	t_{OP}	n_{SEND}	lb_{CO}	lb_{NCO}	t_{DEL}	t_{EXE}	t_{op}/t_{EXE}	ε
Event-driven	396	831	111	2362	527	564	0.702	5.86E-4
Periodical 1	396	10338	118	329	435	472	0.839	5.99E-05
Periodical 3	396	3432	119	596	434	472	0.839	1.82E-05
Periodical 5	396	2040	118	949	429	467	0.848	3.07E-4
Periodical 10	396	1149	112	1362	484	521	0.760	4.63E-4
On-demand	396	2418	123	109	406	444	0.892	2.84E-4

Table 4: Experimental results with random domains of 5 Nodes.

	t_{OP}	n_{SEND}	lb_{CO}	lb_{NCO}	t_{DEL}	t_{EXE}	t_{op}/t_{EXE}	ε
Event-driven	396	1325	117	1451	476	513	0.772	4.26E-4
Periodical 1	396	16880	122	211	426	464	0.853	3.86E-05
Periodical 3	396	5720	122	440	435	471	0.841	1.12E-4
Periodical 5	396	3440	121	640	435	472	0.839	1.84E-4
Periodical 10	396	1840	116	1153	465	525	0.754	2.97E-4
On-demand	396	4050	123	108	415	452	0.876	1.66E-4

Table 5: Experimental results with random domains of 7 nodes.

	t_{OP}	n_{SEND}	lb_{CO}	lb_{NCO}	t_{DEL}	t_{EXE}	t_{op}/t_{EXE}	ε
Event-driven	396	1946	123	137	396	433	0.915	3.61E-4
Periodical 1	396	23198	127	177	418	456	0.868	2.97E-05
Periodical 3	396	7392	124	375	402	439	0.902	9.46E-05
Periodical 5	396	5152	123	595	465	503	0.787	1.18E-4
Periodical 10	396	2632	116	1054	475	512	0.773	2.13E-4
On-demand	396	5523	124	155	396	433	0.915	1.28E-4

Comparing results with the ones achieved with the Global policies, table 2 shows that the execution times are worse, but the exchange of information is reduced in a 25% providing then a better efficiency. This behavior means that the domain-based policies are more scalable. That is, when the number of nodes is increased not only better efficiency results will be obtained, but also better response times.

Random Domains Finally, the results achieved by the policies based on random domains are analyzed. Experiments with domains with 3, 5 and 7 nodes have been done. Regarding

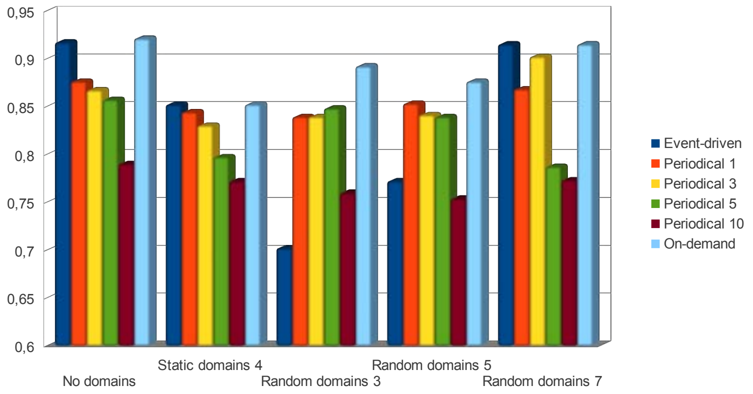


Figure 1: Ratio t_{OP}/t_{EXE}

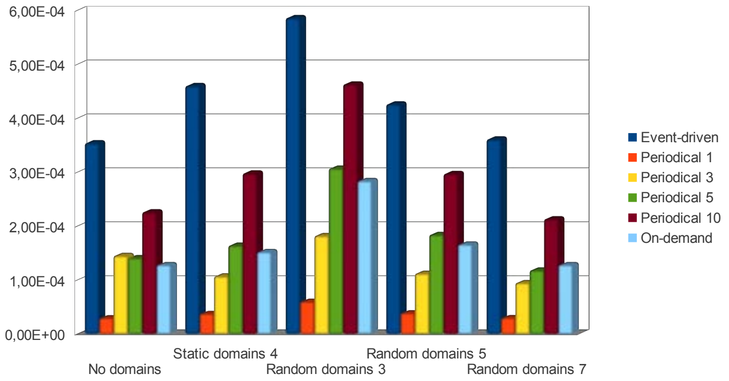


Figure 2: Information efficiency

the times shown in Figure 1, the on-demand policy achieves the best results for all the tested sizes. This is due to the fact that it is quite important to have the most updated information, since the components of the domain are randomly selected. By contrast, the event-driven policy is the one that gives worse results with small domains of 3 and 5 nodes. The reason for this is that sometimes it happens that the state of a node changes to *Recipient* but this information is not sent to the node that is delivering the tasks. This situation can be solved by increasing the domain size. Tables 3, 4 and 5 show an example of this situation. Observing the number of load balancing operations rejected, it is quite high in the case of domains with 3 and 5 nodes. But it can be seen this value is lower in the case of domains with 7 nodes. This value is also significantly reduced in the case of the on-demand policy with respect to the global policies. It can also be seen that the behavior of the periodical policies is similar to the one presented with the global policies.

However, attending to the efficiency values, once again the event-driven policy is the one that achieves the best results, showing better values as smaller is the domain. This is due to the lower number of messages exchanged, with a great difference with the other policies.

4 Conclusions and Future Work

The information exchange rule specifies the way the information about the different nodes workload is collected and maintained, something essential to take the decisions related to the load balancing. One processor should ideally know every time the real state of the rest of the nodes available in the system. However, this is not possible in massively parallel systems, since it would be considerably prejudicial to the algorithm's scalability. So a good information exchange rule has to balance the cost of collecting information and keeping an accurate view of the state of all the nodes in the system.

This paper presents an analysis of the information exchange policies most currently used, as the on-demand, the periodical and the event-driven policies are. Each node stores its information state independently, since this is a distributed algorithm. Keeping a global view of the whole system in massive parallel systems is unfeasible, so local policies must be used. These policies group nodes into domains so the exchange of information takes place only between nodes belonging to the same domain. It preserves the load balancing algorithm scalability, independently of system's size.

This work presents preliminary experiments, since the system used has only a few nodes. Anyway some interesting conclusions can be extracted. As a general conclusion it can be said the event-driven policy is the one that behaves the best on all setups. This is because the response times achieved are quite similar to the ones achieved by the other policies, even better in some cases. But also the number of exchanged messages is quite lower than with the other policies, something that means achieving a better efficiency since the information exchanged is better used. Another consequence is a minimum number of load balancing operation requests rejected.

Future works are focused in two main directions. On one hand, to do this study in a massively parallel system, so the policies based on domains can show their power. On the other hand, to deeply analyze the efficiency metrics, to achieve a better correlation between the efficiency and the execution times of the different experiments.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Education and Science (grants TIN2010-21289, TIN2010-21291-C02-02, Consolider CSD2007-00050 and Cajal Blue Brain project) as well as by the HiPEAC European Network of Excellence.

References

[1] The top500 project. November 2012. http://www.top500.org.

[2] Bestoun S. Ahmed, Khairulmizam Samsudin, Abdul Rahman Ramli, and ShahNor Basri. A descriptive performance model of a load balancing single system image. *Asia International Conference on Modelling and Simulation*, 0:180–184, 2008.

[3] Marta Beltrán and José Luis Bosque. Information policies for load balancing on heterogeneous systems. In *5th International Symposium on Cluster Computing and the Grid (CCGrid 2005), 9-12 May, Cardiff, UK*, pages 970–976. IEEE Computer Society.

[4] J. L. Bosque, O. D. Robles, P. Toharia, and L. Pastor. Load balancing algorithm for heterogeneous clusters. In *12th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2012*, pp. 207–218.

[5] J. L. Bosque, P. Herrero, M. Salvadores, and M. S. Pérez. A collaborative-aware task balancing delivery model for clusters. In *GPC*, pages 146–157, 2007.

[6] N.K. Gondhi and D. Pant. An evolutionary approach for scalable load balancing in cluster computing. In *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pages 1259 –1264, march 2009.

[7] Wenzheng Li and Hongyan Shi. Dynamic load balancing algorithm based on fcfs. In *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, pages 1528 –1531, dec. 2009.

[8] J. Martnez, F. Almeida, E. Garzn, A. Acosta, and V. Blanco. Adaptive load balancing of iterative computation onheterogeneous nondedicated systems. *The Journal of Supercomputing*, 58:385–393, 2011. 10.1007/s11227-011-0595-3.

[9] Xiao Qin, Hong Jiang, Adam Manzanares, Xiaojun Ruan, and Shu Yin. Communication-aware load balancing for parallel applications on clusters. *IEEE Trans. Comput.*, 59(1):42–52, January 2010.

[10] Chengzhong Xu and Francis Lau. *Load Balancing in Parallel Computers: Theory and Practice.* Kluwer Academic Publishers, Boston, 1997.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Symbolic Computation of a Canonical Form for a Class of Linear Functional Systems

Mohamed S. Boudellioua[1]

[1] *Department of Mathematics and Statistics, Sultan Qaboos University, Muscat, Oman.*

emails: boudell@squ.edu.om

Abstract

A matrix pencil is a matrix in the form $\lambda E - A$ where E and A are matrices with elements in a domain D and λ is an indeterminate. If E and A have their elements in say $\mathbb{R}$, then such pencils may arise in the study of systems of differential algebraic equations. In this paper, we consider a more general class of matrix pencils, arising from linear systems of functional equations, where the entries of E and A belong to a polynomial ring e.g. $D = \mathbb{R}[\alpha]$. A canonical form based on the companion matrix and the Smith form is presented for a class of such systems. Examples of these systems are linear neutral delay-differential equations or linear systems of PDEs which are first order with respect to one of the unknown functions. Necessary and sufficient conditions are presented under which a given matrix pencil is equivalent to the canonical form. The exact form of the equivalence transformation is set out using symbolic computations. Examples are also presented to illustrate the ideas developed in this paper.

Key words: Functional systems, Matrix pencils, Companion form, Smith form, Unimodular equivalence.

1 Introduction

Canonical forms play an important role in the modern theory of linear systems. One particular form that has proved to be very useful for 1-D linear systems is the so-called companion matrix which is associated with the characteristic polynomial of a given square matrix. For example, Barnett [1] showed that many of the concepts encountered in 1-D linear systems theory can be nicely linked via the companion matrix. In this paper, we consider a class of multivariate polynomial matrices in the form $\lambda E(\alpha) - A(\alpha)$. This form of matrices arises for example in the study of linear systems of delay-differential equations such as suggested by Byrnes *et al.* [4], where $\lambda = d/dt$ represents a differential operator and α a shift operator, i.e., $\alpha x(t) = x(t-h)$. The matrix $\lambda E(\alpha) - A(\alpha)$, referred to as a matrix pencil can be regarded as a generalization of the matrix pencil $\lambda E - A$ often encountered in the study of linear systems of differential algebraic equations. A canonical form associated with the characteristic polynomial of the pair $E(\alpha), A(\alpha)$ is defined for these pencils. Necessary and sufficient conditions are presented under which a matrix pencil is equivalent to resulting canonical form. The exact form of the transformation connecting the pencils is established. First we present a few definitions that will be needed later in the paper.

Definition 1 *Let $D = K[x_1, \ldots, x_n]$. The general linear group $GL_p(D)$ is defined by*

$$GL_p(D) = \left\{ M \in D^{p\times p} \;\mid\; \exists N \in D^{p\times p} : MN = NM = I_p \right\} \tag{1}$$

An element $M \in GL_p(D)$ is called a unimodular matrix. It follows that M is unimodular if and only if the determinant of M is invertible in D, i.e., is a non-zero element of K.

One of standard tasks carried out in systems theory, is to transform a given system representation into a simpler form before applying any analytical or numerical method. The transformation involved must of course preserve relevant system properties if conclusions about the reduced system are to remain valid about the original one. An equivalence transformation used in the context of linear functional systems is unimodular equivalence. This transformation can be regarded as an extension of Rosenbrock's equivalence [11] from the univariate to the multivariate setting and is defined by the following.

Definition 2 *Let T_1 and T_2 denote two $q\times p$ matrices with elements in D then T_1 and T_2 are said to be unimodular equivalent if there exist two matrices $M \in GL_q(D)$ and $N \in GL_p(D)$ such that*

$$T_2 = MT_1N \tag{2}$$

2 Reduction using Unimodular Equivalence

The reduction of multivariate polynomial matrices arising from linear functional systems was studied by a number of authors see for example, Frost and Storey [8], Lee and Zak [9]. In particular, Frost and Boudellioua [7] gave necessary and sufficient conditions under which a class of bivariate polynomial matrix is unimodular equivalent to a Smith form corresponding to a simpler system containing only one equation in one unknown. Lin *et al.* [10] extended this result to the multivariate case. Boudellioua and Quadrat [3] generalized this result using a module theoretic approach.

Theorem 1 ([7]) *Let $T \in D^{n\times n}$, with full row rank, then T is unimodular equivalent to the Smith form*

$$S = \begin{pmatrix} I_{n-1} & 0 \\ 0 & |T| \end{pmatrix} \tag{3}$$

if and only if there exist a vector $U \in D^n$ which admits a left inverse over D such that the matrix $\begin{pmatrix} T & U \end{pmatrix}$ has a right inverse over D.

3 Matrix Pencils over $D = \mathbb{R}[\lambda, \alpha]$

In this section we shall consider a type of matrix in the form $\lambda E(\alpha) - A(\alpha)$. Suppose that in Theorem 1, $D = \mathbb{R}[\lambda, \alpha]$ and

$$T = \lambda E(\alpha) - A(\alpha) \tag{4}$$

where $E(\alpha)$ and $A(\alpha)$ are $n \times n$ matrices with elements in $\mathbb{R}[\alpha]$ and suppose that there exists a vector $U \in D^n$ which admits a right inverse over D. Then it follows that the matrix T is equivalent to the Smith form:

$$S = \begin{pmatrix} I_{n-1} & 0 \\ 0 & |\lambda E(\alpha) - A(\alpha)| \end{pmatrix} \tag{5}$$

where $|\lambda E(\alpha) - A(\alpha)|$ is the characteristic polynomial associated with the matrix pair $(E(\alpha), A(\alpha))$.
The following interesting result follows.

Lemma 1 (Canonical form) *Consider a matrix T in the form (4) for which the condition in Theorem 1 holds and let*

$$|T| \equiv |\lambda E(\alpha) - A(\alpha)| = \sum_{i=0}^{n} e_i(\alpha)\lambda^{n-i}, \qquad (e_0(\alpha) \textit{ monic})$$

Then the matrix pencil T is equivalent to the canonical form:

$$\overline{T} = \lambda S_E(\alpha) - F(\alpha) \tag{6}$$

where $S_E(\alpha)$ is the Smith form:

$$S_E(\alpha) = \begin{pmatrix} I_{n-1} & 0 \\ 0 & e_0(\alpha) \end{pmatrix}$$

and $F(\alpha)$ is the $n \times n$ companion matrix:

$$F(\alpha) = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -e_n(\alpha) & -e_{n-1}(\alpha) & \cdots & -e_1(\alpha) \end{pmatrix}$$

References

[1] S. Barnett *A matrix circle in linear control theory*, I.M.A. Bulletin **12** (1976) 173–176.

[2] M.S. Boudellioua *Computation of the Smith form for multivariate polynomial matrices using Maple*, American J. of Computational Mathematics **2** (2012) 21–26.

[3] M.S. Boudellioua and A. Quadrat *Serre's reduction of linear functional systems*, Mathematics in Computer Science **4** (2010) 289–312.

[4] C.I Byrnes, M.W. Spong, and T.J. Tarn *A several complex variables approach to feedback stabilization of linear neutral delay-differential systems*, Mathematical Systems Theory **17** (1984) 97–133.

[5] F. Chyzak, A. Quadrat, and D. Robertz OreModules*: A symbolic package for the study of multidimensional linear systems*, In J. Chiasson and J.-J. Loiseau, editors, Applications of Time-Delay Systems, Springer, `http://wwwb.math.rwth-aachen.de/OreModules/` **LNCIS 352** (2007) 233–264.

[6] A. Fabianska and A. Quadrat *Applications of the Quillen-Suslin theorem in multidimensional systems theory*, In H. Park and G. Regensburger, Editors, *Grobner Bases in Control Theory and Signal Processing, Radon Series on Computation and Applied Mathematics, de Gruyter Publisher* **3** (2007) 23–106.

[7] M.G. Frost and M.S. Boudellioua *Some further results concerning matrices with elements in a polynomial ring*, Int. J. Control **43** (1986) 1543–1555.

[8] M.G. Frost and C. Storey *Equivalence of a matrix over* $\mathbb{R}[s, z]$ *with its Smith form*, Int. J. Control **28** (1979) 665–671.

[9] E.B. Lee and S.H. Zak *Smith forms over* $\mathbb{R}[z_1, z_2]$, IEEE Trans. Autom. Control **28** (1983) 115–118.

[10] Z. Lin, M.S. Boudellioua and L. Xu *On the Equivalence and Factorization of multivariate polynomial matrices*, Proceedings of the 2006 international symposium of circuits and systems, Island of Kos (Greece) (2006).

[11] H. H. Rosenbrock, *State space and multivariable theory*, Nelson-Wiley, London (1970).

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Efficient protocols to control glioma growth

J. R. Branco[1], J. A. Ferreira[2] and P. de Oliveira[2]

[1] *CMUC & Department of Physics and Mathematics, Coimbra Institute of Engineering, 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 consider a mathematical model to describe glioma evolution. The model is established combining the viscoelastic behaviour of the brain tissue with a mass conservation law that takes into account the effect of chemotherapy. For the non Fickian model we establish an upper bound for the tumor mass that leads to a sufficient condition to control tumor growth. Based on the theoretical upper bound, protocol for chemotherapy treatment are proposed. Numerical experiments are included to illustrate the behaviour of the model as well as the efficiency of the presented protocols.

Key words: Viscoelastic behaviour, tumor growth, glioma, chemotherapy.

1 Introduction

Gliomas are the most common type of brain tumors. They begin in the glial cells and thus diffuse and highly invade the brain tissue, often intermixing with normal brain tissue. Unfortunately, the prognosis for patients with gliomas is very poor. Median untreated survival time for high grade gliomas ranges from 6 months to 1 year and even lower grade gliomas can rarely be cured. Tumor cell transport and proliferation are the main contributors to the malignant dissemination [21]. Theorists and experimentalists believe that inefficiency of treatments results of the highly mobility capacity and high proliferation rates presented by glioma cells.

Research activity in the mathematical modelling of tumor growth has been very fruitful specially in solid tumors where the growth primarily comes from cellular proliferation. Glioma's growth is characterized by proliferation (as solid tumors) but also by invasion of the surrounding brain tissue. The recognition that tumor cells might spread outside the grossly visible mass, invading locally and metastasizing distantly, and that some cells die during the development process, lead to more complex mathematical concepts than those used in the original simple models for solid tumors ([8], [11], [12], [14], [18], [20], [21]).

The most popular model used to measure the glioma growth is characterized by an equation of type

$$\frac{\partial c}{\partial t} = \nabla(D\nabla c) + f(c)\,, \text{ in } \Omega \times (0, +\infty)\,. \tag{1}$$

where $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3,$ is the spatial domain of the glioma, $c(x, t)$ denotes the tumor cell density at location x and time t, $f(c)$ denotes net proliferation of tumor cells, D represents the diffusion tensor and ∇ defines the spatial gradient operator (see [18]). The proliferation term f is assumed to be exponential, and so cell growth term is given by $f(c) = \rho\, c$, where the net proliferation rate ρ is constant. Logistic and gompertzian growths are also possible choices for f but found to be unnecessary in the time frames considered for gliomas [14].

Equation (1) is established combining the mass conservation law

$$\frac{\partial c}{\partial t} + \nabla J_F = f(c)\,, \text{ in } \Omega \times (0, +\infty), \tag{2}$$

with the classical Fick's law for the mass flux J_F,

$$J_F = -D\nabla c\,. \tag{3}$$

The partial differential equation (1) is of parabolic type and it is well known that if a sudden change on the cell concentration takes place somewhere in the space, it will be felt instantaneously everywhere. This means that Fickian approach gives rise to infinite speed of propagation which is not a physical property. To avoid this limitation of Fickian models an hyperbolic correction has been proposed in different contexts ([1], [2], [6], [7], [13], [16], and the references cited therein).

In this paper we consider a mathematical model to describe glioma growth of non Fickian type that takes into account the viscoelastic behaviour of the brain tissue ([10], [15] and [17]). Following [2], [3], [4], [5] and [19], the viscoelastic behaviour of the brain tissue is included in the definition of the mass flux considering the effect of the stress exerted by the brain tissue on the tumor cells.

Chemotherapy is one of the most popular treatments used on gliomas. This therapy involves the use of drugs to disrupt the cell cycle and to block proliferation. The success of chemotherapy agents varies widely, depending on cell type and the type of drug being used. The effectiveness of a particular drug is dependent on the concentration of drug reaching the tumor, the duration of exposure and the sensitivity of the tumor cells to the drug.

Tracqui *et al.* [22] incorporated chemotherapy by introducing cell death as a loss term. If $G(t)$ defines the time profile of the chemotherapy treatments then, assuming a loss proportional to the amount of therapy at a given time, equation (1) is replaced by

$$\frac{\partial c}{\partial t} = \nabla(D\nabla c) + f(c) - G(t)c\,, \text{ in } \Omega \times (0, +\infty)\,, \tag{4}$$

where

$$G(t) = \begin{cases} k, & \text{when chemotherapy is being administered} \\ 0, & \text{otherwise}\,. \end{cases} \tag{5}$$

Here k describes the rate of cell death due to exposure to the drug. If $f(c) = \rho c$, for a tumor to decrease in size during chemotherapy, k must be larger than the growth rate ρ of the cell population. The main question is to define k and the periods of chemotherapy applications that lead to control the glioma mass.

The mathematical model that we consider is defined in a simple geometry. To apply the modeling approach to specific patients, a more realistic look at the brain geometry and structure was necessary. In [20] Swanson *et al.* introduced the complex geometry of the brain and allowed diffusion to be a function of the spatial variable x to reflect the observation that glioma cells exhibit higher motility in the white matter than in grey matter ([11]).

The paper is organized as follow. In Section 2 we present a class of non Fickian models that describe the space and time evolution of glioma cells constructed combining the diffusion process with the viscoelastic properties of the brain tissue. In Section 3 we study the behaviour of the glioma mass and we establish sufficient conditions on the parameters of the model that lead to control glioma growth. These sufficient conditions allow us to define the standard bang-bang chemotherapy protocol. In Section 4 we present numerical experiments that illustrate the effect of several protocols. Finally, in Section 5 we include some conclusions.

2 A viscoelastic model

In this section we present the mathematical model that will be considered in this work. Following [2], [3], [4], [5] and [19], if a diffusion process occurs in a medium that has a viscoelastic behaviour then this behaviour should be included in the mass flux. This fact means that the mass flux J admits the representation

$$J = J_F + J_{nF}, \tag{6}$$

where the Fickian flux J_F is given by (3) and the non Fickian mass flux J_{nF} is defined by

$$J_{nF}(t) = -D_v \nabla\sigma(t), \tag{7}$$

where σ represents the stress exerted by the brain tissue on the tumor cells.

We will assume that the viscoelastic behaviour of the brain tissue is described by

$$\frac{\partial \sigma}{\partial t} + \beta \sigma = \alpha_1 \epsilon + \alpha_2 \frac{\partial \epsilon}{\partial t}, \tag{8}$$

where ϵ stands for the strain. Equation (8) is based on a mechanistic model which is represented by a spring (restorative force component) and a dashpot (damping component) in parallel connected with a free spring. In (8) the viscoelastic characteristic time β is given by $\beta = \frac{E_0+E_1}{\mu_1}$, and $\alpha_1 = \frac{E_0 E_1}{\mu_1}$, $\alpha_2 = E_0$, 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 for instance [10], [15] and [17]).

Equation (8) leads to the following expression for σ

$$\sigma(t) = \int_0^t e^{-\beta(t-s)} (\alpha_1 \epsilon(s) + \alpha_2 \frac{\partial \epsilon}{\partial t}(s)) ds + e^{-\beta t} \sigma(0). \tag{9}$$

If we assume that the strain ϵ satisfies $\epsilon = \lambda c$ where λ is a positive constant (see [2], [3], [4] and [5]) from (9) we obtain

$$\sigma(t) = \lambda \int_0^t e^{-\beta(t-s)} (\alpha_1 c(s) + \alpha_2 \frac{\partial c}{\partial t}(s)) ds + e^{-\beta t} \sigma(0). \tag{10}$$

Mass conservation equation (2) with J_F replaced by J, given by (6), leads to the integro-differential equation

$$\frac{\partial c}{\partial t} = \nabla(D^* \nabla c) + \int_0^t k_{er}(t-s) \nabla(D_v^* \nabla c(s)) + f(c)\,, \text{ in } \Omega \times (0, +\infty)\,, \tag{11}$$

where $D^* = D + \lambda \alpha_2 D_v\,,\ D_v^* = \lambda(\alpha_1 + \alpha_2) D_v$ and $k_{er}(t) = e^{-\beta t}\,.$

To establish a mathematical model to describe the evolution in time and space of the glioma cells some medical information is needed. According to [8] and [9] the following assumptions are assumed in our model:

- glioma cells are of two phenotypes: proliferative (state 1) and migratory (state 2);
- in state 1 cells randomly move but there is no cell fission;
- in state 2 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 switches to a cell of type 2;
- β_1 is the switching rate from state 1 to state 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 state 1.

Let $u(x,t)$ and $v(x,t)$ be the densities of migratory and proliferation cells at position x and time t, respectively. The dynamics of glioma cells in $\Omega \times (0,T]$ is then described by (11), where we have dropped the asterisk in D^* and D_v^*, completed with an equation that describes the dynamic of proliferation cells

$$\left\{\begin{array}{l} \dfrac{\partial u}{\partial t} = \nabla(D\nabla u) + \displaystyle\int_0^t k_{er}(t-s)\nabla(D_v \nabla u(s)) - \beta_1 u + \beta_2 v\,, \\ \dfrac{\partial v}{\partial t} = \rho v + \beta_1 u - \beta_2 v\,, \end{array}\right. \tag{12}$$

where D and D_v denote square matrices of order n, β_1 is the switching rate from migratory phenotype to proliferative phenotype and β_2 is the switching rate from proliferative phenotype to migratory phenotype.

If chemotherapy is applied and $G(t)$ defines the time profile of the chemotherapy treatments then, assuming a loss proportional to the amount of therapy at a given time, system (12) is replaced by

$$\left\{\begin{array}{l} \dfrac{\partial u}{\partial t} = \nabla(D\nabla u) + \displaystyle\int_0^t e^{-\beta(t-s)}\nabla(D_v \nabla u(s)) - \beta_1 u + \beta_2 v - G(t)u\,, \\ \dfrac{\partial v}{\partial t} = \rho v + \beta_1 u - \beta_2 v - G(t)v\,, \end{array}\right. \tag{13}$$

where $G(t)$ is defined by (5).

System (13) is completed with initial conditions

$$u(0) = u_0, \;\; v(0) = v_0 \text{ in } \Omega, \tag{14}$$

and boundary conditions

$$u(t) = v(t) = 0 \text{ on } \partial\Omega, \tag{15}$$

where $\partial\Omega$ denotes the boundary for Ω. Condition (15) means that glioma is located inside the brain and cancer cells do not attain pia mater.

3 Control of glioma growth

We will assume that $D = [d_{ij}]$ and $D_v = [d_{v,ij}]$ are diagonal matrices such that

$$0 < \alpha_e \leq d_{ii}, d_{v,ii} \leq \alpha_b \text{ in } \overline{\Omega},\; i = 1, \ldots, n. \tag{16}$$

Let $\mathcal{M}_1(t)$ be the natural total mass of tumor cells in Ω,

$$\mathcal{M}_1(t) = \int_\Omega (u(t) + v(t))\, d\Omega \tag{17}$$

and $\mathcal{M}_2(t)$ be an artificial mass of tumor cells, defined by the accumulated energy

$$\mathcal{M}_2(t) = \|u(t)\|^2 + \|v(t)\|^2 \,, \tag{18}$$

where $\|.\|$ denotes the usual L^2 norm which is induced by the usual L^2 inner product $(.,.)$. For mathematical reasons we will study the behaviour of the artificial mass of tumor cells $\mathcal{M}_2(t)$, hoping to control the natural total mass $\mathcal{M}_1(t)$.

We have

$$\frac{1}{2}\mathcal{M}_2'(t) = \int_\Omega \left(\frac{\partial u}{\partial t}(t)u(t) + \frac{\partial v}{\partial t}(t)v(t) \right) \, d\Omega.$$

From (13) and taking into account the boundary conditions (15) we deduce

$$\begin{aligned}\frac{1}{2}\mathcal{M}_2'(t) \;=\; & -\|\sqrt{D}\,\nabla u(t)\|^2 - \Big(\int_0^t e^{-\beta(t-s)} D_v \nabla u(s)\, ds, \nabla u(t)\Big) + (\,-\beta_1 - G(t))\|u(t)\|^2 \\ & +(\rho - \beta_2 - G(t))\|v(t)\|^2 + (\beta_1 + \beta_2)(u(t), v(t)),\end{aligned} \tag{19}$$

where $\sqrt{D} = [\sqrt{d_{ii}}\,]$.
As

$$\begin{aligned}\Big(\int_0^t e^{-\beta(t-s)} D_v \nabla u(s)\, ds, \nabla u(t)\Big) \;=\; & \frac{1}{2}\frac{d}{dt}\|\int_0^t e^{-\beta(t-s)}\sqrt{D_v}\nabla u(s)\, ds\|^2 \\ & +\beta\,\|\int_0^t e^{-\beta(t-s)}\sqrt{D_v}\,\nabla u(s)\, ds\|^2,\end{aligned}$$

and

$$\alpha_e\|v\|^2 \leq C_\Omega^2\|\sqrt{D}\,\nabla v\|^2, \;\; v \in H_0^1(\Omega), \tag{20}$$

then from (19) we get

$$\begin{aligned}\frac{d}{dt}\Big(\mathcal{M}_2(t) + \|\int_0^t e^{-\beta(t-s)}\sqrt{D_v}\nabla u(s)\, ds\|^2\Big) \leq & -2\beta\|\int_0^t e^{-\beta(t-s)}\sqrt{D_v}\nabla u(s)\, ds\|^2 \\ & +2\max\Big\{\frac{\beta_2-\beta_1}{2} - \frac{\alpha_e}{C_\Omega^2} - G(t), \frac{\beta_1-\beta_2}{2} + \rho - G(t)\Big\}\mathcal{M}_2(t).\end{aligned} \tag{21}$$

If

$$\frac{\beta_1-\beta_2}{2} + \rho > \frac{\beta_2-\beta_1}{2} - \frac{\alpha_e}{C_\Omega^2} \tag{22}$$

and

$$\frac{\beta_2-\beta_1}{2} - \frac{\alpha_e}{C_\Omega^2} - G(t) > -\beta, \tag{23}$$

then equation (21) leads to

$$\mathcal{M}_2(t) \leq e^{2\left((\frac{\beta_1-\beta_2}{2}+\rho)t - \int_0^t G(s)\, ds\right)}\mathcal{M}_2(0). \tag{24}$$

As $\dfrac{\alpha_e}{C_\Omega^2}$ is a constant arising from mathematical analysis, conditions (22), (23) can be replaced by

$$\beta_1 - \beta_2 + \rho > 0$$

and

$$\beta_1 - \beta_2 < 2(\beta - G(t)),$$

respectively. Avoiding such constant, these last conditions assume a biological meaning. To conclude that the artificial mass $\mathcal{M}_2(t)$ is bounded by $\mathcal{M}_2(0)$, we need to combine conditions (22), (23) with

$$\Big(\frac{\beta_1 - \beta_2}{2} + \rho\Big)t < \int_0^t G(s)\,ds. \tag{25}$$

Condition (25) means that density of proliferation cells at time t, that is density of cells originated by cells of this type and cells that comes from state 1 and remains in state 2, is less than the total amount of death cells until time t due to chemotherapy effect.

From Schwarz inequality we have

$$\mathcal{M}_1(t) \le \sqrt{|\Omega|}\,(\|u(t)\| + \|v(t)\|). \tag{26}$$

If we assume that $\sqrt{|\Omega|} \le \|u(t)\|$ and $\sqrt{|\Omega|} \le \|v(t)\|$, then we conclude that the upper bound (24) for $\mathcal{M}_2(t)$ is also an upper bound for the mass $\mathcal{M}_1(t)$. We note that inequality (26) has pure mathematical character and it is not obviously that it has a medical translation. However for the different simulations that we carried on, inequality (26) was verified and consequently we can use condition (25) to control tumoral mass.

When chemotherapy is applied, condition (25) 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, provided that condition (22) holds. 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 (5) and illustrated in Figure 1.

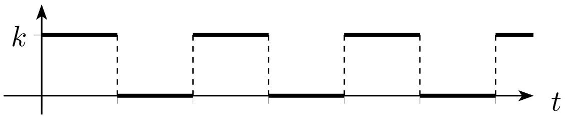


Figure 1: Chemotherapy protocol.

4 Numerical simulation

In this section we present some numerical results illustrating the behaviour of the glioma cells defined by (13). The numerical results were obtained using a standard numerical method defined combining the explicit Euler methods with second order centered difference operators and a rectangular rule to discretize the spatial derivatives and the time integral, respectively. We consider a homogeneous square domain $\Omega = [0, 15\,cm] \times [0, 15\,cm]$, diffusion coefficients $d_{11} = d_{22} = d_{v,11} = d_{v,22} = 0.025\,cm^2/day$, growth rate $\rho = 0.05\,/day$, switching parameters $\beta_1 = 10^{-6}/day$ and $\beta_2 = 0.036/day$, kernel such that $\beta = 1$ and initial condition defined by 10^6 proliferation tumor cells at middle square $[7, 8] \times [7, 8]$.

In Figure 2 we plot the numerical solutions at day 33 for an virtual untreated patient $(G(t) = 0)$. We observe a decreasing on the highest values of the tumor cells concentration at initial times followed by an increase and very intense spreading of cells. The contour plots allow us to observe high gradients on the core of the tumor, defined by the proliferation cells, and that the migration cells are already quite far from the core!

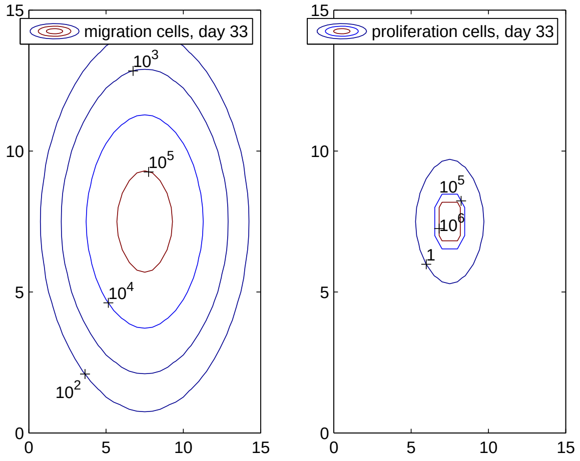


Figure 2: Numerical results at day 33, obtained with 2D model (13) for $k = 0/day$.

Let us consider now that the chemotherapy treatment defined by (5) is applied with a protocol as illustrated in Figure 1. Condition (25) is used to compute an effective drug 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 allowed by condition (25), for a virtual patient as defined in the beginning os this section. Here $\alpha_e = 0.025\,cm^2/day$ and $C_\Omega = \frac{1}{\sqrt{2}}$.

Protocol	$k_{\min}$ [./day]
each 2 days	0.064
each 7 days	0.224
each 14 days	0.448

Table 1: $k_{\min}$ as (25), for a protocol of 24 consecutive hours of chemotherapy .

In Figure 3 we plot the cell distribution at day 33, when a protocol of chemotherapy of 24h is administered at days 5, 19 and 33 using $k = 0.5/day$. We observe that glioma mass at day 33 is less than its mass at day 4 (the day before the first administration of the chemotherapy).

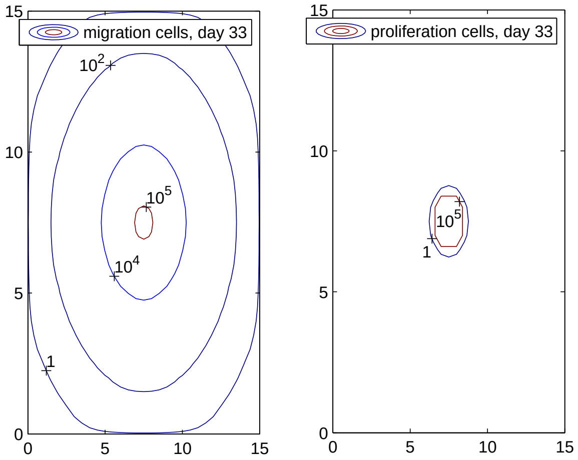


Figure 3: Numerical results at day 33, obtained with 2D model (13) for $k = 0.5/day$.

Finally, in Figure 4 we compare glioma masses of the virtual patient when no chemotherapy is administered and the results of the adminstration of 3 different chemotherapy protocols. The difference between the protocols is the rest period and the values of k were computed using condition (25). We observe that for all protocols glioma masses are less than the glioma mass at day 4 (the day before the first administration of the protocol). The results presented in this figure shows the effectiveness of the our approach to define chemotherapy protocols.

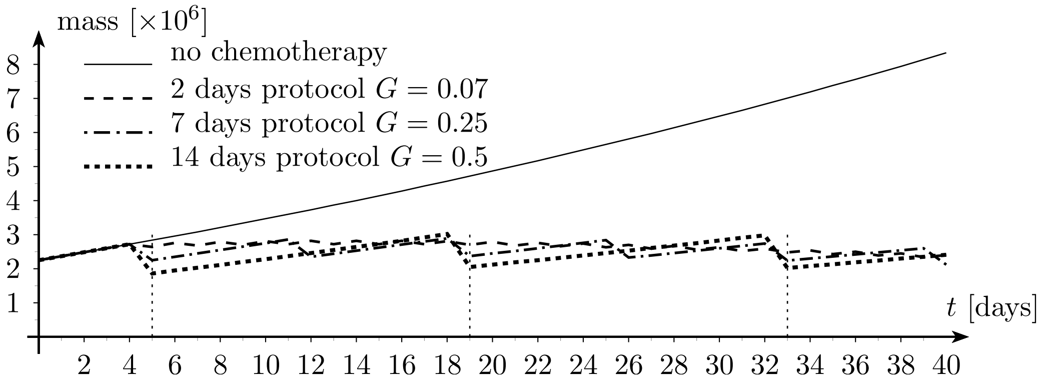


Figure 4: Glioma masses $\mathcal{M}_1(t)$.

5 Conclusions

In this paper we studied a mathematical model to describe the evolution of glioma cells when chemotherapy is applied. The model was established combining a mass conservation with a non Fickian mass flux that takes into account the viscoelastic behaviour of the brain tissue described by the Voigt-Kelvin model.

Using the energy method we deduced an estimate for the glioma mass $\mathcal{M}_2(t)$, defined using L^2 norm. This estimate allowed us to define a sufficient condition on the parameters of the model that leads to the control of $\mathcal{M}_2(t)$, more precisely, to guarantee that $\mathcal{M}_2(t) < \mathcal{M}_2(0)$. Such condition was then used to define chemotherapy protocols. Numerical experiments illustrating the behaviour of the glioma mass under the conditions deduced for the chemotherapy protocols are also included. The results obtained suggest our approach is a promising one. Future work will address the comparison of the model with existing medical protocols.

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 Fundação para a Ciência e Tecnologia (FCT) under the projects PEstC/MAT/UI0324/2011 and by the project UTAustin/MAT/0066/2008.

References

[1] J. R. Branco, J. A. Ferreira, P. de Oliveira, *Numerical methods for the generalized Fisher-Kolmogorov-Petrovskii-Piskunov equation*, Applied Numerical Mathematics, **57** (2007) 89–102.

[2] D. A. Edwards, D. S. Cohen *An unusual moving boundary condition arising in anomalous diffusion problems*, SIAM Journal on Applied Mathematics, **55** (1995) 662–676.

[3] D. A. Edwards, D. S. Cohen *A mathematical model for a dissolving polymer*, AIChE Journal, **41** (1995) 2345–2355.

[4] D. A. Edwards, *Non-Fickian diffusion in thin polymer films*, Journal of Polymer Science, Part B: Polymer Physics Edition, **34** (1996) 981–997.

[5] D. A. Edwards, *A spatially nonlocal model for polymer-penetrant diffusion*, Journal of Applied Mathematics and Physics, **52** (2001) 254–288.

[6] S. Fedotov, *Traveling waves in a reaction-diffusion system: diffusion with finite velocity and Kolmogorov-Petrovskii-Piskunov kinetics*, Physical Review E, **58** (1998) 4:5143–5145.

[7] S. Fedotov, *Nonuniform reaction rate distribution for the generalized Fisher equation: Ignition ahead of the reaction front*, Physical Review E, **60** (1999) 4:4958–4961.

[8] S. Fedotov, A. Iomin, *Migration and proliferation dichotomy in tumor-cell invasion*, Physical Review Letters, **98** (2007) 118110(1)–(4).

[9] S. Fedotov, A. Iomin, *Probabilistic approach to a proliferation and migration dichotomy in tumor cell invasion*, Physical Review E, **77** (2008) 1031911(1)–(10).

[10] G. Franceschini, *The mechanics of human brain tissue*, PhD thesis, University of Trento, 2006.

[11] A. Giese, L. Kluwe, B. Laube, H. Meissner, M. Berens, M. Westphal, *Migration of human glioma cells on myelin*, Neurosurgery, **38** (1996) 755–764.

[12] S. Habib, C. Molina-París, T. Deisboeck, *Complex dynamics of tumors: modeling an energing brain tumor system with coupled reaction-diffusion equations*, Physica A, **327** (2003) 501–524.

[13] S. Hassanizadeh, *On the transient non-Fickian dispersion theory*, Transport in Porous Media, **23** (1996) 107–124.

[14] H. L. Harpold, E. C. Alvord Jr, K. R. Swanson, *The evolution of mathematical modeling of glioma proliferation and invasion*, rm Journal of Neuropathology and Experimental Neurology, **66** (2007) 1–9.

[15] J. Humphrey, *Continuum biomechanics of soft biological tissues*, rm Proceedings of Royal Society London, **459** (2003) 3–46.

[16] D. Joseph, L. Preziosi, *Heat waves*, Review of Modern Physics, **61** (1989) 47–71.

[17] A. Mehrabian, Y. Abousleiman, *A general solution to poroviscoelastic model of hydrocephalic human brain tissue*, Journal of Theretical Biology, **29** (2011) 105–118.

[18] J. D. Murray, *Mathematical Biology*, Springer, 2002.

[19] 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.

[20] K. R. Swanson, E.C. Alvord Jr, J.D. Murray, *A quantitative model for differential motility of gliomas in grey and white matter*, Cell Proliferation, **33** (2000) 317–329.

[21] K. R. Swanson, C. Bridge, J. D. Murray, E. C. Alvord Jr, *Virtual and real brain tumors: using mathematical modelig to quantify glioma growth and invasion*, Journal of the Neurological Sciences, **216** (2003) 1–10.

[22] 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 Proliferation, **28** (1995) 17–31.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Allee effects models in randomly varying environments

Carlos A. Braumann[1] and Clara Carlos[2]

[1] *Centro de Investigação em Matemática e Aplicações, Universidade de Évora*

[2] *Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal*

emails: braumann@uevora.pt, clara.carlos@estbarreiro.ips.pt

Abstract

Based on a deterministic model of population growth with Allee effects, we propose a general stochastic model that incorporates environmental random fluctuations in the growth process. We study the model properties, existence and uniqueness of solution and the stationary behavior. We also obtain expressions for the first passage times, in particular, the mean and standard deviation of extinction times for the population.

Key words: Allee effects, population growth, random environments, extinction times

1 Introduction

Warder Clyde Allee (1885/1955) was an American zoologist and ecologist who taught animal ecology. The Allee effects were first described by Allee and colleagues in 1949, as we can see in [2]. There are two kinds of Allee effects, strong Allee effects and weak Allee effects. When a population has a "critical size or density" below which the population decreases on average and above which it increases on average, it is called a strong Allee effect. On the other hand, when a population does not exhibit a "critical size or density" but at low densities the population growth rate increases with increasing density, we say there is a weak Allee effect. Allee effects show up in many wildlife populations, particularly when low population size hinders the efficacy of collective defensive behavior from predators or results in individuals being far apart, which makes it difficult to find mating partners.

We introduce a quite general deterministic model of population growth with strong Allee effect, particular cases of which can be seen in [6], [8], [10] or [1]. The stochastic model we propose is based on this deterministic model with an added term to account for the effect of environmental fluctuations on the growth rate. Other stochastic Allee effects models have been proposed in [7] and [9] but they consider the effect of demographic stochasticity (stochasticty due to sampling variations in births and deaths in an unchanging environment). The model proposed here is important to understand the consequences of environmental variability affecting the growth rate. We study the properties of the proposed stochastic model.

We also characterize the time the population takes to reach a given size for the first time. In particular, we study the extinction time for the population.

2 Model

Let $X = X(t)$ be the population size at time $t > 0$ and L a positive constant. We propose as a general model with strong Allee effects that the *per capita* growth rate be of the form

$$\frac{1}{X}\frac{dX}{dt} = f(X), \tag{1}$$

where $f(X)$ is a real C^1 function defined for $X > 0$ such that $-\infty < f(0^+) < 0 < f(L)$, $f(+\infty) < 0$ and, strictly, $f(X)$ increases for $X < L$ and decreases for $X > L$. We assume that the initial size $X(0) = x$ is known.

In [3] we can find such models with no Allee effects, in particular the classical logistic model, corresponding to $f(X) = r\left(1 - \frac{X}{K}\right)$, with intrinsic growth rate $r > 0$ and carrying capacity of the environment $K > 0$. A model of population growth with strong Allee effects similar to the logistic models is $f(X) = r\left(1 - \frac{X}{K}\right)\left(\frac{X}{E} - 1\right)$, with $0 < E < K$ (see, for instance, [6]). An Allee limit E, with $0 < E < L < K$ such that $f(E) = 0$, is incorporated such that the *per capita* population growth is negative below E.

In a randomly fluctuating environment, the *per capita* growth rate varies randomly and expression (1) should now be interpreted as describing its average behavior and, since growth is a multiplicative type process, the geometric average is the appropriate one to consider. We then need to add the effect of environmental fluctuations on the *per capita* growth rate, which we assume to be of the white noise type, of the form $\sigma\epsilon(t)$, where $\epsilon(t)$ is a standard continuous-time white noise and $\sigma > 0$ is the noise intensity, assumed to be constant and independent of population size. Therefore, assuming expression (1) represents the geometric average growth rate, it can be seen in [5] that the appropriate stochastic calculus to use is the Stratonovich calculus, and so we shall use it here. We obtain the stochastic differential equation (SDE)

$$\frac{1}{X}\frac{dX}{dt} = f(X) + \sigma\epsilon(t), \tag{2}$$

with $X(0) = x$ known.

The solution $X(t)$ exists and is unique up to an explosion time. We can show that there is no explosion and therefore, the solution exists and is unique for all $t \geq 0$. The solution

$X(t)$ is a homogeneous diffusion process with drift coefficient $a(x) = x\left(f(x) + \frac{\sigma^2}{2}\right)$ and diffusion coefficient $b^2(x) = \sigma^2 x^2$. Let us define, in the interior of the state space, the scale and speed measures of $X(t)$. The scale density is

$$s(y) := \exp\left(-\int_n^y \frac{2a(\theta)}{b^2(\theta)} d\theta\right) = \frac{n}{y} \exp\left(-\frac{2}{\sigma^2} \int_n^y \frac{f(\theta)}{\theta} d\theta\right) \tag{3}$$

and the speed density is

$$m(y) := \frac{1}{b^2(y)s(y)} = \frac{1}{n\sigma^2 y} \exp\left(\frac{2}{\sigma^2} \int_n^y \frac{f(\theta)}{\theta} d\theta\right), \tag{4}$$

where n is an arbitrary (but fixed) point in the interior of the state space and $C > 0$ is a constant. The corresponding "distribution" functions are $S(z) = \int_c^z s(y)dy$ and $M(z) = \int_c^z m(y)dy$, the scale function and speed function, respectively, where c is an arbitrary (but fixed) point in the interior of the state space. The scale measure is define by $S(a,b) = S(b) - S(a)$ and the speed measure is by $M(a,b) = M(b) - M(a)$.

The state space has boundaries $X = 0$ and $X = +\infty$. One can see that $X = 0$ is attracting but unattainable, so there is no t such that $X(t) = 0$; however, it may happen that $X(t) \to 0$ when $t \to +\infty$ and so "mathematical" extinction can occur. One can see that $X = +\infty$ is non-attracting and, therefore, explosions can not occur.

Contrary to the deterministic model (1), the stochastic model (2) does not have an equilibrium point, but it may exist an equilibrium probability distribution for the population size, called the stationary distribution, with a probability density function $p(y)$, known as stationary density. The stationary density $p(y)$, when it exists, must satisfy the Kolmogorov forward equation

$$\frac{d(a(y)p(y))}{dy} - \frac{1}{2}\frac{d^2(b^2(y)p(y))}{d^2y} = 0, \tag{5}$$

We can show that every non-negative solution is not integrable $\left(\int_0^{+\infty} p(y)dy = +\infty\right)$ and, consequently, contrary to the corresponding stochastic model without Allee effects, there is no stationary density.

In fact, "mathematical" extinction does occur, since $X(t) \to 0$ when $t \to +\infty$, due to the attractiveness of 0 and non-attractiveness of $+\infty$. But we prefer to use the concept of "realistic" extinction, meaning the population dropping below an extinction threshold $q > 0$ (for example, $q = 1$). We are interested in the extinction time, i.e. the time $T_q = \inf\{t > 0 : X(t) = q\}$ required for the population to reach the extinction threshold for the first time. Assuming $0 < q < x < +\infty$, we can write (see, for instance, [4]) expressions for the n-th order moment of T_q :

$$M_q^{(n)}(x) = E\left[(T_q)^n | X(0) = x)\right] = 2\int_q^x s(\zeta) \int_\zeta^{+\infty} n\, M_q^{(n-1)}(\theta) m(\theta) d\theta d\zeta. \tag{6}$$

Since $M_q^{(0)}(x) = 1$, one can iteratively obtain the moments of any arbitrary order of T_q and, of course, also the mean and the variance.

Acknowledgements

The authors are members of the Centro de Investigação em Matemática e Aplicações, a research center of the Universidade de Évora financed by FCT (Fundação para a Ciência e Tecnologia, Portugal). Carlos A. Braumann gratefully acknowledges the financial support from the FCT grant PTDC/MAT/115268/2009-Dynamics of epidemiological spreading and the FAPESP Thematic Project-Assessing control of epidemics using mathematical and computer models.

References

[1] P. Amarasekare, *Allee effects in metapopulation dynamics*, The American Naturalist, Vol. 152, n.2, 298–302 (1998).

[2] W. C. Allee, A. E. Emerson, O. Park , T. Park and K. P. Schmidt, *Principles of Animal Ecology.* Saunders, Philadelphia (1949).

[3] C. A. Braumann, *Applications of stochastic differential equations to population growth*, Proceedings of the Ninth International Colloquium on Differential Equations (Bainov, D., Ed.), VSP, Utrecht, p. 47–52 (1999).

[4] C. A. Braumann, P. A. Filipe, C. Carlos and C. J. Roquete, *Growth of individuals in randomly fluctuating environments*, Proceedings of the International Conference in Computational and Mathematical Methods in Science e Engineering, Vigo-Aguiar, J., Alonso, P., Oharu, S., Venturino, E. and Wade, B. (Eds.), Gijon, 201–212 (2009).

[5] C. A. Braumann, *Itô versus Stratonovich calculus in random population growth*, Mathematical Biosciences, **206**, 81–107 (2007).

[6] F. Courchamp, T. H. Cliutton-Brock and B. Grenffell, *Inverse density dependence and the Allee effect*, Trends in Ecology & Evolution, **14**: 405–410 (1999).

[7] B. Dennis, *Allee effects in stochastic populations*, Oikos, **96**: 389–401 (2002).

[8] B. Dennis, *Allee effects: population growth, critical density and the change of extinction*, Natural Resource Modelling, **3**: 481–539 (1989).

[9] S. Engen, R. Lande and B.-E. Sæther, *Demographic stochasticity and Allee effects in populations with two sexes*, Ecology, **84(9)**, 2378-2386 (2003).

[10] M. A. Lewis and P. Kareiva, *Allee dynamics and the spread of invading organisms*, Theoretical Population Biology **43**: 141–158 (1993).

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Exploiting Multi-core Platforms for Multi-model Forest Fire Spread Prediction

Carlos Brun[1], Tomàs Margalef[1] and Ana Cortés[1]

[1] *Computer Architecture and Operating Systems Department*
Universitat Autònoma de Barcelona

emails: carlos.brun@caos.uab.es, tomas.margalef@uab.es, ana.cortes@uab.es

Abstract

The Two-Stage forest fire spread prediction methodology was developed to enhance forest fire evolution forecast by tackling the uncertainty of certain environmental conditions. There are parameters, such as wind, that present a variation along terrain and time. In such cases, it is necessary to develop multi-model prediction schemes that integrate forest fire propagation models and complementary models, such as meteorological forecast and wind field models. This multi-model approach should improve the accuracy of the predictions, but introducing an overhead in the execution time. In this paper, different multi-model approaches are discussed and the results show that the propagation prediction is improved. The overhead introduced by the complementary models can be overlapped with the data observations so the final prediction time is not increased significantly.

Key words: HPC, forest fire, prediction, multi-core, efficiency, multi-model.

1 Introduction

Forest fires is a worrisome hazard that every year causes important losses around the world. For that reason, there has been a great research activity in this field during the last decades, in order to develop models which try to represent and predict the behavior of such hazard [1][2]. Forest fire spread simulators [3][4] require a set of input parameters describing the environmental conditions, the initial fire front, the topography of the terrain and the vegetation. However, this input data far from being easy to obtain, it arises as one of the main problems to tackle. Topographical data is the most reliable input of a simulator, however, the remaining parameters suffer from different degrees of uncertainty ranging from the total ignorance to very low resolution. For that reason, prediction schemes based on executions of hundreds or even thousands of fire scenarios (different simulator input configurations) with the aim of filling the gap of uncertainty have been proposed [5]. Two-Stage Prediction strategy [6] is a prediction scheme, which performs a forest fire prediction by previously executing a Calibration stage which involves the most sensitive parameters such as environmental conditions. In this Calibration stage, the actual evolution of the forest fire is observed and a Genetic Algorithm (GA) is carried out to determine the set of parameters that best reproduces the recent evolution of the fire. This set of values is then used as input parameters in the Prediction stage. As it is well stated, using a GA as a tunning strategy will imply a cost in terms of execution time, depending on the time spent for evaluating each individual. In our case, one individual implies running once FARSITE (the underlying forest fire spread simulator) and, since one single fire spread simulation can last from seconds to almost one hour, the time incurred in providing a useful prediction could become prohibitive. Furthermore, the original Two-Stage prediction scheme suffers from two main handicaps. This scheme considers an uniform distribution of the parameters along the whole terrain and it does not consider prognostic models to enable dynamic parameters changes through time. Both restrictions have a direct impact in the quality of the prediction results [4][7]. Thus, the original scheme was modified to be a multi-model prediction framework where different complementary models were easy to couple, in order to reduce that negative impact. Therefore, we focus on the meteorological conditions and, in particular, in the wind components, since these are the parameters that most affect fire spread [8]. In order to consider the meteorological wind modifications due to the topography, a wind field model must be introduced to obtain the effective wind at the required level of detail. On the other hand, to enable the two-stage prediction scheme the capacity of reacting to sudden changes in environmental conditions, it becomes mandatory to fit into the prediction scheme, environmental data coming from prognostic models such as weather forecasting models. Both prognostic models and wind field models are computationally expensive. So, any approach to couple those models into a system (Two-Stage prediction scheme in our case) will need to carefully analyze the implications in the total execution time and resources needed. When we deal with natural hazards such as forest fires, any fire spread prediction must be delivered faster than real time fire evolution to be useful. In this paper, we propose a multi-model prediction framework for forest fire spread prediction. This framework involves different models such as forest fire spread model, wind field model and a meteorological model. Due to the computational needs required for this framework, we rely on High Performance Computing (HPC). Furthermore, since predicting the evolution of forest fires implies working under very tight real time constraints, we also analyzed how to use the available computing resources in the most efficient way without loosing prediction quality.

In the next section, the proposed multi-model prediction framework is discussed. In section 3 the characteristics of the experiment performed and the execution platform used are introduced. In section 4, a study of the prediction quality provided when different complementary models are included into the framework is performed. The execution time and efficiency of each approach is studied in section 5 and, finally the main conclusions of this work are reported in section 6.

2 Multi-model forest fire propagation prediction framework

In the above mentioned Two-Stage prediction scheme, the calibrated input parameter set is determined applying a GA. A random population of individuals, each one representing a scenario, is generated. Each individual is simulated and the fire propagation obtained is compared to the real fire propagation. According to the quality of the prediction, the individuals are ranked and the genetic operators are applied to generate the new population. The process is repeated a certain number of iterations and the best individual at the end of the process is selected to run the prediction for the next time interval. The GA fits very well in the Master/Worker paradigm. The Master process generates the initial random population. Then, it distributes the individuals to the Worker processes that can be executed independently on different cores. Each core in the platform can run a Worker process. So, the maximum number of Worker processes is limited to the number of available cores in the platform. If the population is larger than the number of cores, then some individuals must be executed sequentially. Once the Workers have executed the FARSITE simulation corresponding to its individual, they evaluate the error compared to the actual fire propagation using the symmetric difference among the real and simulated burned area. The error obtained is returned to the Master process that ranks the individuals and applies the genetic operators to generate the next population. This process is repeated for a fixed number of iterations. This scheme has been called Two-Stage basic prediction methodology (2ST-BASIC prediction) and it is shown in figure 1(a). In this scheme, if there are enough available cores, the execution time of each iteration is limited by the execution time of the scenario whose simulation lasts longer. The quality of the calibration depends on the available elapsed time to provide the propagation prediction and the number of individuals on each iteration, but it must be considered that in these emergency situations, response time is a critical issue [9]. For the previous studies [10] it was stated that 5 iterations usually provide a successful calibration.

2.1 Coupling Wind Field Model to 2ST-BASIC (2ST-WF)

As it has been mentioned above, some parameters present a spatial and temporal distribution which makes the calibration process more difficult since only an average value for the parameter along map and time can be selected. One of such parameters is wind. The wind can be measured by meteorological stations, but the value measured in one meteorological station is a measure in a single point, but in other points of the terrain the value (speed and direction) of the wind can be others due to the orography of the terrain. So, it is necessary to introduce some complementary model that calculates the wind speed and direction in any point of the terrain given a meteorological wind. The selected wind field model is WindNinja [7] because it has a direct connection to FARSITE. In this scheme, each worker process receives the parameters representing one particular scenario and then, it is necessary to run the WindNinja wind modeler and afterwards the FARSITE fire propagation simulator. This approach is more realistic but the computational time required by WindNinja is quite large. The execution time of the WindNinja modeler depends on the map size and topography, but usually it takes around some minutes on a single core. This pipelined worker scheme is depicted in figure 1(b).

2.2 Coupling Meteorological Model to 2ST-BASIC (2ST-MM)

It is well known that wind can change suddenly in speed and direction. During the calibration stage, it is feasible to receive information from meteorological stations frequently (every 30 minutes or even more frequently). In this case, the wind speed and direction do not need to be calibrated since they are received from direct measurements. However, during the prediction stage such values are not available beforehand. So, it is necessary to introduce a meteorological model that can provide the expected values for the meteorological wind speed and direction. This values can be used during the prediction stage [11]. In this work, we assume that meteorological predictions are available from a meteorological service, so, it is not necessary to compute the meteorological forecast on the fire spread prediction platform. It implies that the computational cost of the 2ST-BASIC forest fire propagation prediction (see figure 1(c)) is not increased.

2.3 Complete Multi-model forest fire propagation scheme (2ST-WF-MM)

Next step is integrating all the mentioned models on a multi-model forest fire propagation framework (figure 1(d)). In this case, the meteorological wind speed and direction are not calibrated, but they are provided by meteorological model and then introduced to the wind field model to provide the wind field. In this case, the meteorological measures and the meteorological predictions are known before the prediction stage starts and, therefore, the wind field corresponding to each time interval can also be computed beforehand.

3 Experiment design

For the experimental part of this work, we chose a relevant area in Catalonia (north-east of Spain) as the landscape to be simulated. In particular, we concentrated on the north-east

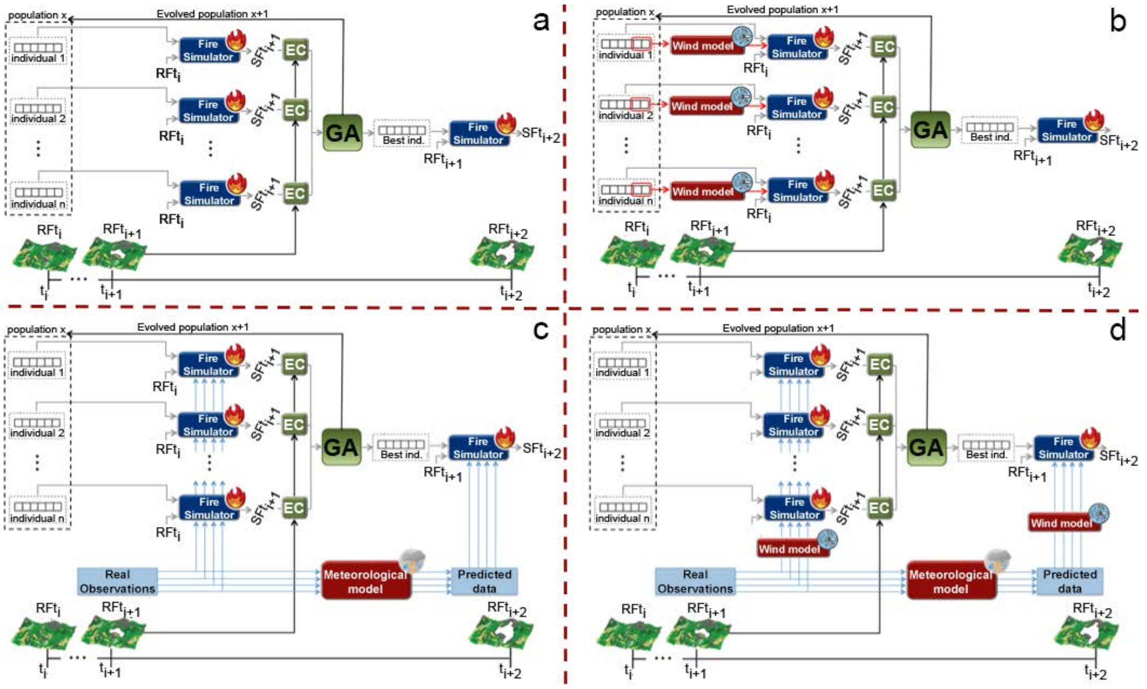


Figure 1: 2ST-BASIC (a), 2ST-WF (b), 2ST-MM (a) and 2ST-WF-MM (b) multi-model prediction schemes

cape (*El Cap de Creus*) with an approximate real extension of 900 square kilometers. In order to evaluate all proposed schemes, we created a reference fire that lasts 12 hours. The components of the global meteorological wind (wind speed and wind direction) vary every 30 minutes including significant variations between stages. The resulting fire evolution is stored and used as a real fire evolution, and the input settings that were used to generate this propagation are dismissed. Although this, for this test case, the meteorological data has been generated in a synthetic way. We have already evaluated the viability to obtain in real time this information from the corresponding local weather forecast service (*Servei Metereologic de Catalunya*) [12][13]. This information involves both kind of data needed, real observations provided by meteorological weather stations and forecasted data delivered by the WRF (Weather Research and Forecasting) model. The time window selected for the Calibration and Prediction stages was chosen taking into account the time period needed to gather useful information from weather forecast services and satellite sensor systems. In the former case, the typical time-step of coarse scale weather forecast models ranges from 3 to 6 hours, what determines the frequency of delivered data. On the later case, we should consider the time interval required for receiving fire front images that could properly be used in that multi-model prediction framework. To obtain such a perimeter information, we rely on sensors systems which are on board both the NASA's Terra and Aqua satellites. It is necessary for each satellite to complete 3 orbits (approximately 3 hours) to cover the whole Europe area, so it could be possible to obtain fire perimeters every 3-6 hours [14]. Consequently, the time window selected for the Calibration stage and the Prediction stage, in this test case, has been 6 hours each. Therefore, we had 12 hours to execute the whole prediction scheme independently on the multi-model scheme selected. Since the Calibration strategy implements a GA, we perform the experiment for different population sizes such as 25, 50 and 100 individuals in order to analyze the influence of this parameter on the results in terms of quality, execution time and efficiency. The GA has been iterated 5 times and, for each initial population size, we have performed 5 different experiments with different initial population. Thus, the results reported in the following sections are the mean values of those 5 experiments.

Finally, the execution platform used for all the results reported in the following sections is an IBM cluster x3650 composed of 32 nodes with two Dual-Core Intel(R) Xeon(R) CPU 5150 2.66GHz and 12 GB Fully Buffered DIMM 667 MHz each.

4 Quality analysis

The schemes described in section 2 incorporate a GA that iteratively improves the quality of the calibration. So, it is necessary to analyze the convergence of the GA and the influence of other parameters such as the population size in the quality delivered of every multi-model scheme. Figure 2 summarizes the evolution of the error at each generation for all the 4 schemes. It can be observed that the schemes which incorporate the meteorological model (2ST-MM and 2ST-MM-WF) do not require a significant number of generations to achieve an almost stable calibration error. The main reason of this behavior is that, in these cases, the wind is not a parameter to be calibrated, but it is a measured or forecasted parameter. Therefore, the calibration process is easier since the wind speed and direction are very relevant parameters that must be calibrated carefully. In this cases, the population size does not appear to be a relevant factor since the errors are very similar. The schemes that do not incorporate meteorological model (2ST-BASIC and 2ST-WF) must calibrate the meteorological wind value and, therefore, the calibration process requires more iterations. In this schemes, the population size is a relevant factor since larger populations provide better calibration results. It is also worthy observing that the schemes that incorporate the wind field model (2ST-WF and 2ST-MM-WF)) provide better results than those that use a general value for the whole terrain. It means that the wind field model is an added value to the prediction process. The calibration process is very significant, but finally the most relevant result is the prediction error. Figure 3(a) shows the errors obtained at the end of the calibration process and the corresponding prediction errors. The same information is depicted in figures 3(b) where a visualization of the delivered predicted fire front evolution for each multi-model scheme is plotted. As we can observe, the multi-model scheme that provides better results both in the Calibration stage as in the Prediction stage, is the 2ST-MM-WF as it was expected. Let's analyze the results scheme by scheme in more detail. The 2ST-BASIC is the scheme that delivers worst results both in prediction and calibration errors. This scheme needs to perform all the iterations of the GA to reduce the prediction results but, even iterating until the preset number of generations, it is not able to adapt to changes in the meteorological conditions such as wind parameters. The initial population size could slightly reduce this penalty but, in general terms, this effect is independent on the number of the GA's individuals. When observing the results provided by the 2ST-WF, we detected a quality improvement in the Calibration stage. This enhancement is due to, not only the inclusion of the wind field evaluation for each particular individual, but also to the ability of calibrating the general wind components. This ability enables the system to better adjust the wind parameters to reproduce more precisely the recent past behavior of the fire. However, this enhancement is not extrapolated to the Prediction stage. As we can observe, the prediction error drastically increases despite having a good calibration error. The main reason of this is that wind values do not keep quite constant from Calibration stage to Prediction stage. Thus, 2ST-WF scheme is not able to adapt to meteorological changes, if those changes happens during the Prediction interval. As it is was reported in section 2, to overcame this drawback, a meteorological model was coupled into the multi-model system. The advantages of including such a model are reflected in the prediction errors obtained for 2ST-MM and 2ST-MM-WF. Using forecasted data in the Prediction stage helps the system to dynamically adapts to changes in environmental variables. However, 2ST-MM delivers worse results than 2ST-MM-WF. The reason of this difference is the ability of 2ST-MM-WF to provide high resolution wind flow that better reproduces the wind variations at a field level. So, if the wind speed varies because of a mountain or a valley the 2ST-MM-WF multi-model scheme captures this effects. Those results keep quite similar for all initial population sizes, so, a last conclusion in this point could be that, a population size of 25 individuals is enough to obtain a reasonable prediction results. However, these conclusions must be contrasted with the execution time and efficiency results reported in the next section.

5 Execution Time and Efficiency

Prediction quality is a very important issue, but execution time is as critical as accuracy. So, it is necessary to study the execution time of each scheme and the efficiency reached. The execution time of each scheme for different population sizes (25, 50 and 100 individuals) have been evaluated and the results are shown in table 1. As it can be observed, the execution time of the 2ST-WF is by far, the most time consuming multi-model scheme. The need of executing a wind field model for each individual at each iteration results in a

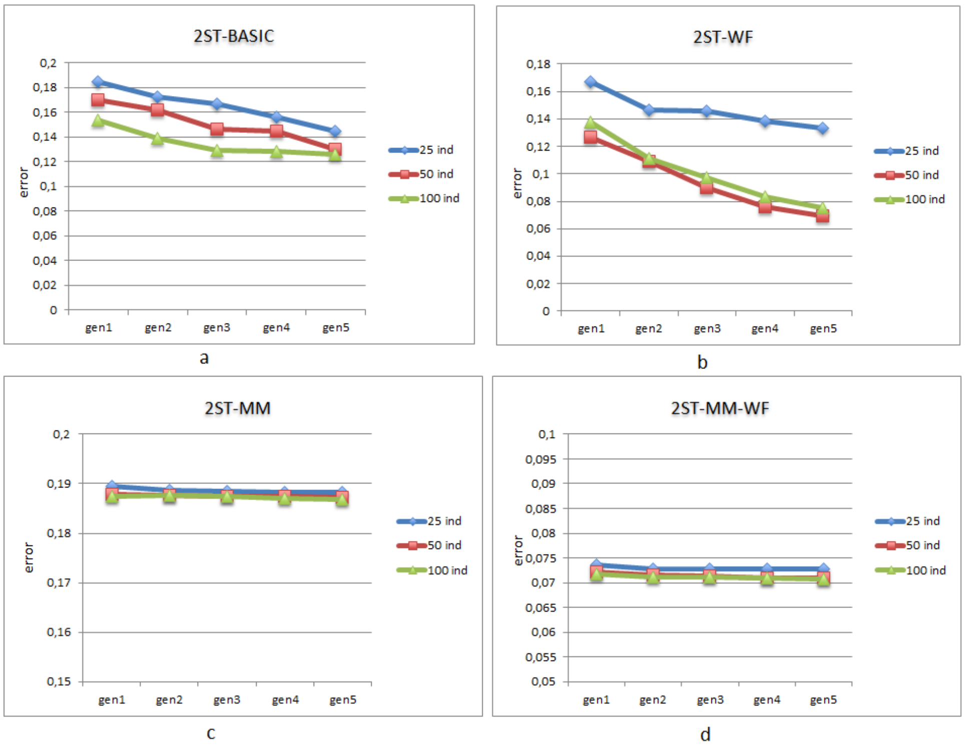
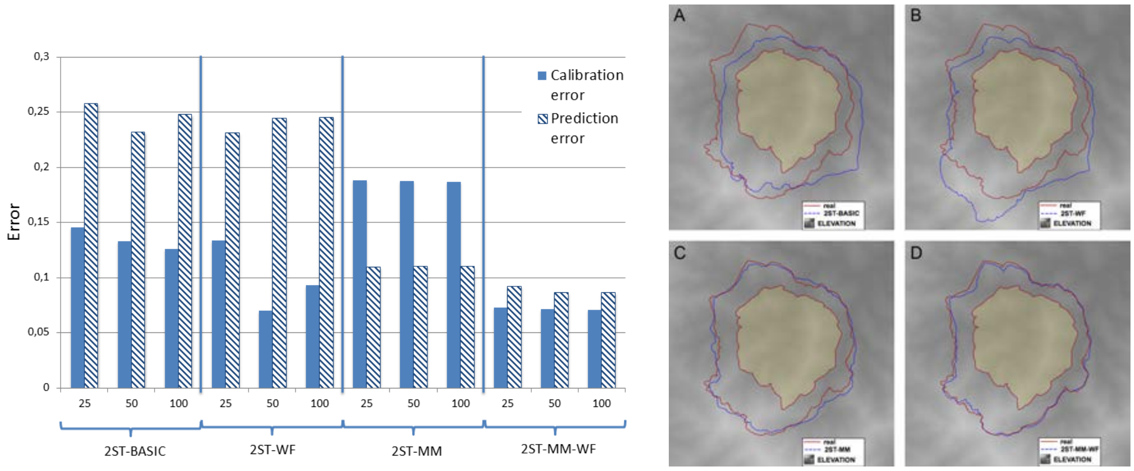


Figure 2: Prediction error evolution for 2ST-BASIC (a), 2ST-BASIC-WF (b), 2ST-BASIC-MM (c) and 2ST-BASIC-MM-WF (d) using initial populations of 25, 50 and 100 individual

increment of time. So, for the shake of understanding, figure 4, summarizes the execution time of all schemes except 2ST-WF. The execution time depends on the number of workers, the number of nodes on the system, the architectural features of each node, and more specifically on the multi-model scheme selected and the particular scenario represented. As it was described in section 3, for running our test we use 128 cores. Since our populations are 100 individuals at most, there are enough cores to execute one worker per core, but the internal architecture of each node has a crucial effect. Each node has 2 processors and 12GB of memory. Internally each processor has 2 cores with a local cache memory of 64KB (32KB for data and 32KB for instructions) per core and a shared cache of 4MB. The forest fire propagation simulator needs information concerning the features of the terrain where the fire is taking place with a significant resolution. It means that they must usually manage around 3MB of data. Actually, the data size depends on the map size and the resolution, but this is a typical size. Therefore, when we can allocate one worker per node or even per processor, each worker has the 4MB of shared cache memory available and this amount of memory is enough to store all the required data. However, when the number of workers



(a) (b)

Figure 3: Calibration and Prediction errors for all multi-model predictions schemes using initial populations of 25, 50 and 100 individuals (a) and the predicted fire front when 25 individuals are used (b).

Population size	**25**	**50**	**100**
2ST-BASIC	31.38s	41.16s	70.27s
2ST-WF	829.16s	1083.45s	1609.38s
2ST-MM	34.06s	47.25s	74.07s
2ST-MM-WF	63.71s	73.72s	97.97s

Table 1: Execution time of every scheme and population size.

increases and two workers are allocated in the two cores of the same processor, they must share the 4MB of cache memory. This memory is to small to host the data corresponding to both workers and then they must continuously access to main memory with the significant execution time degradation. Another factor that affects the execution time when the number of individuals is increased, is the access to the data files. 100 workers simultaneously reading the input data files (terrain, vegetation, etc.) and writing the propagation maps in files in NFS, introduce a bottleneck that contributes to increase the execution time. So, larger populations (100 individuals) need less generations to converge to a reasonable error, but the execution time of each generation is significantly larger than the shorter populations (25 or 50 individuals). So, the use of more resources does not provide the expected benefit in execution time. Comparing the execution time obtained by the different schemes, it can be observed that the 2ST-BASIC is the fastest scheme. It was expected since it is the scheme

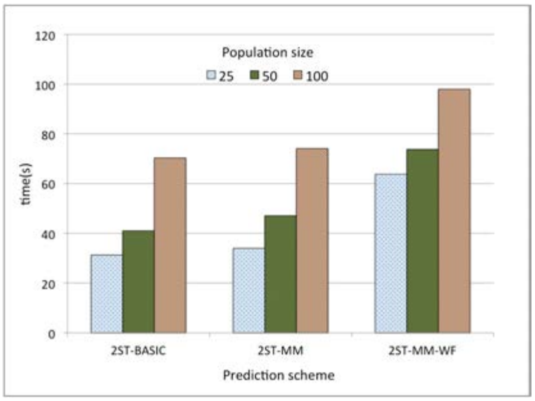


Figure 4: Execution time of 2ST-BASIC, 2ST-MM, 2ST-MM-WF for all population sizes

that involves less models (actually only forest fire simulation is involved) and moreover, all the parameters are uniform and constant. However, it can be seen that it needs several iterations of the GA to converge and with the same number of iterations the calibration and prediction errors are higher than the ones obtained using the other schemes. The 2ST-WF scheme presents an execution time that, in its current state, cannot be assumed. The main problem incurred by this scheme is that it is necessary to evaluate the wind field corresponding to every individual in any iteration, and this enlarges the duration of a single iteration. The average execution time of the WindNinja in the tested scenario is around 220 seconds. Considering that the number of iterations is 5, then the time incurred by the wind field calculation along the 5 generations is around 1100 seconds. Moreover, the execution time of FARSITE when a wind field is introduced is larger than the FARSITE execution time when a uniform wind is considered. To make this scheme viable, it is necessary to accelerate the execution of the wind field model. The 2ST-MM scheme assumes that the measurements of the meteorological stations and the forecasted data provided by meteorological models, are available before the prediction stage is launched. So, this scheme does not introduce any additional computational cost. Instead, since the wind is not introduced as one of the parameters to be calibrated, the calibration process is much shorter and the required number of iterations is significantly reduced, with the corresponding reduction in prediction time. The complete multi-model 2ST-MM-WF scheme does not require to calibrate the wind parameters since they are measured or predicted, and then, the calibration requires less iterations. So, the computational cost of introducing the Wind field modeler does not suppose a significant time increase. The only overhead introduced is the one incurred when evaluating the wind field in the Prediction Stage.

6 Conclusions

Fire spreading is a complex phenomena that implies, not only, the execution of a fire spread model, but also complementary models, which enrich the description of the environmental conditions where the fire takes place. In particular, it is crucial to be able to determine the wind components at a high resolution for the whole terrain. So, meteorological models that forecast the wind components at a low resolution and wind fields models, which are able to move this low resolution winds to a high resolution wind flows at a field level, are required. We have proposed and analyzed three different multi-models fire spread prediction schemes, which are based on the so called Two-Stage prediction method. The inclusion of these models have a direct impact in the total execution time of the prediction scheme, as well as in the number of computing resources needed to run them. So, the performance of the proposed multi-model schemes has been analyzed in terms of prediction quality improvements, execution time incurred to deliver a prediction and their efficiency. From the experimental study outcomes that coupling both models, meteorological and wind field models, to the forest fire spread model improves the quality of the predictions. However, in order to use the resources in a more efficient way, the GA implemented in the calibration stage of the proposed methodology achieves reasonable calibration errors with populations sizes about 25 individuals. That means that it is not necessary to use large populations, instead, the resources released as a consequence of shorten the population, could be directly applied to run the commented complementary models.

Acknowledgments

This research has been supported by MICINN-Spain TIN2011-28689-C02-01.

References

[1] D. X. Viegas. Fire line rotation as a mechanism for fire spread on a uniform slope. *International Journal of Wildland Fire*, 11(1):11–23, 2002.

[2] R.C. Rothermel. *How to predict the spread and intensity of forest and range fires.* US Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station Ogden, UT, USA, 1983.

[3] 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.

[4] A. M. G. Lopes, M. G. Cruz, and D. X. Viegas. Firestation an integrated software system for the numerical simulation of fire spread on complex topography. *Environmental Modelling & Software*, 17(3):269–285, 2002.

[5] G. Bianchini, M. Denham, A. Cortés, T. Margalef, and E. Luque. Wildland fire growth prediction method based on multiple overlapping solution. *J. Comput. Science*, 1(4):229–237, 2010.

[6] B. Abdalhaq, A. Cortés, T. Margalef, and E. Luque. Enhancing wildland fire prediction on cluster systems applying evolutionary optimization techniques. *FGCS The International Journal of Grid Computing*, 21(1):61 – 67, 2005. Methods Applications.

[7] J. M. Forthofer. *Modeling wind in complex terrain for use in fire spread prediction.* PhD thesis, Colorado State University, 2007.

[8] B. Abdalhaq, A. Cortés, T. Margalef, and E. Luque. Evolutionary optimization techniques on computational grids. In *International Conference on Computational Science (1)*, pages 513–522, 2002.

[9] A. Cencerrado, R. Rodríguez, A. Cortés, and T. Margalef. Urgency versus accuracy: Dynamic driven application system natural hazard management. *International Journal of Numerical analysis and Modeling*, (9):432–448, 2012.

[10] M. Denham, K. Wendt, G. Bianchini, A., and T. Margalef. Dynamic data-driven genetic algorithm for forest fire spread prediction. *Journal of Computational Science*, 3(5):398 – 404, 2012.

[11] C. Brun, T. Artés, T. Margalef, and A. Cortés. Coupling wind dynamics into a dddas forest fire propagation prediction system. *Procedia Computer Science*, 9(0):1110 – 1118, 2012.

[12] METEO.CAT (http://www.meteo.cat). Servei meteorologic de catalunya.

[13] A. Cencerrado, M. A. Senar, and A. Cortés. Support for urgent computing based on resource virtualization. In *ICCS (1)*, pages 227–236, 2009.

[14] D. Rodriguez-Aseretto, D. de Rigoa, M. Di Leoa, A. Cortés, and J. San-Miguel-Ayanza. A data-driven model for large wildfire behaviour prediction in europe. *Procedia Computer Science (to be published)*, 2013.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Do niches help in controlling disease spread in ecoepidemic models?

Iulia Martina Bulai[1], **Bruna Chialva**[1], **Davide Duma**[1] **and Ezio Venturino**[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino,*
via Carlo Alberto 10, 10123 Torino, Italy

emails: `iuliam@live.it`, `bruna.chiara@alice.it`, `davideduma@gmail.com`, `email:` `ezio.venturino@unito.it`

Abstract

We present three models for refuges in interacting population systems of predator-prey type, with the prey hosting a transmissible disease. The safety niche is assumed to lessen the disease spread, but not to protect prey from predators. This represents a novelty with respect to standard ecosystems where the refuge prevents predators' attacks. The niche is assumed either to protect the healthy individuals, or to hinder the infected ones to get in contact with the susceptibles, or finally to reduce altogether contacts that might lead to new cases of the infection. Some counterintuitive results are obtained. The effectiveness of the three different strategies are compared. The best situation in terms of disease containment appears to be the environment which provides a place where the healthy individuals cannot come in contact with disease carriers.

Key words: refuge, niches, disease transmission, ecoepidemics
MSC 2000: AMS codes 92D30, 92D25, 92D40

1 Introduction

In population models predator-prey and competition systems play a dominant role, since the blossoming of this discipline about a century ago. In more recent times, more refined models try to better describe reality. Since prey try to seek protection against attacks of their predators in the features of the environment, scientists have tried to incorporate this behavior into the interaction models. The introduction of refuges has lead to the observation that the Lotka-Volterra models gets stabilized [3] even to show global asymptotic stability, [1, 2]. This shows the relevant role that spatial refuges exert in shaping the dynamics of predator-prey interplay. The refuge is expressed in the equations by reducing the amount of prey population available for hunting by the predators.

In this classical setting, if Y denotes the prey population that can take cover, by Y_n we denote the number of individuals who find protection in the niches that are available for their safety. Thus there are only $Y - Y_n$ individuals that can interact with the predators. There could be several functional forms that can be chosen for Y_n. The simplest one is a constant value, $Y_n = Y_0$, with $Y_0 \in \mathbf{R}_+$, or alternatively one could take a linear function of the prey population, $Y_n = Y_0 Y$, [3] or also a linear function of the predators X, $Y_n = Y_0 X$ [6]. More recently, a model has been proposed in which the form is taken as a bilinear function of both populations, $Y_n = Y_0 XY$, [4].

Ecoepidemiology investigates the influence of diseases in ecosystems, see Chapter 7 of [5]. It appears therefore that the refuges for some of the populations involved can be introduced also in this context. However, instead of using the environmental niches as protection against the predators, i.e. as an ecological tool as described above, we employ them in order to investigate whether they can influence the disease spread, i.e. we give them an epidemiological meaning. Therefore, it is not against predators that prey are protected, but we rather consider the case in which the healthy prey for some reason due to the conformation of the environment can avoid to come in contact with disease-carriers of their own population and therefore be somewhat protected from the epidemics. This is achieved by reduced contact rates that they have with infected individuals. Of all the various possible types of niche, to keep things simple, we just take the constant case, $Y_n = Y_0$.

In the next Sections, we present three models, based on the ecoepidemic system presented in [7], differing in the way the refuge is modeled. In Section 2, some of the susceptibles are prevented from interaction with infected individuals. In Section 3, it is part of the infected that are unable to become in contact with healthy individuals. In Section 4, we look at a reduced contact rate. A final discussion compares the results.

2 The model with a refuge for the healthy prey

Consider at first the system in which the susceptibles are more able to wander about than the infected ones, because the latter indeed are in general weakened by the disease. In this way, it is possible that the susceptibles reach places unattainable by the diseased individuals. Thus the latter cannot come in contact with the healthy remote individuals, and therefore these sound individuals cannot be infected. We assume that s denotes the fixed number of susceptibles that escape from the spread of the epidemics using the refuge.

The model is formulated as follows. The healthy prey R reproduce with net reproduction rate a, are subject to intraspecific competition only with other sound individuals at rate b and are hunted by predators at rate c. Those that can be infected by the diseased prey individuals U, as discussed above, leave their class at rate λ, to enter into the class of sick inviduals. The latter do not reproduce, are hunted at a rate $k \neq c$ by the predators. Here $k > c$ means that they are weaker than sound ones, and therefore more easy to capture, while $k < c$ instead takes into account the fact that they might be less palatable than the healthy ones. Finally, they can recover the disease at rate ω and therefore reenter into the S population. As mentioned above, infected are assumed not to contribute to intraspecific pressure, either of sound prey or among themselves; this again is grounded in the fact that their disease-related weakness prevents them to compete with the other individuals in the population. The predators are assumed to have also other food sources, for which they reproduce at rate d, but clearly get a benefit from the interactions with the healthy prey expressed by the parameter $e < c$. This constraint expresses the fact that the amount of food they get from the captured prey cannot exceed its mass. So far all the system parameters are nonnegative. For the predators hunting the infected prey, instead, we could model two different situations. For $h > 0$, the infected cause a damage to the predators, killing them. In this paper we concentrate only on this case. In the opposite case we could have the normal situation in which predators get a reward from capturing the diseased prey, so that in this situation we would have $0 < -h < k$. In summary, the ecoepidemic model with inclusion of a disease-safety niche for the susceptibles reads

$$\begin{aligned} \frac{dR}{dt} &= R[a - bR - cF] - \lambda(R - s)U + \omega U \qquad (1)\\ \frac{dU}{dt} &= \lambda(R - s)U - U[kF + \omega]\\ \frac{dF}{dt} &= F[d + eR - fF - hU] \end{aligned}$$

Note that the above system needs some further qualifications. In fact when $R < s$ the next to last term in the first equation and the first one in the second equation would become positive and negative respectively, which makes no sense biologically. Therefore in such situations they should be understood to be identically zero. But in such case the infected prey in the system are easily seen to vanish, since in the second equation the term on the right hand side is always negative. The system then would settle to one of the equilibria of the classical disease-free predator-prey model, with logistic correction for the prey alternative food supply for the predators, see [7] for its brief analysis. For the benefit of the reader a short summary of its findings is presented also here at the top of Section 5.

The equilibria of (1) are $P_1 = (0, 0, 0)$ and

$$P_2 = \left(0, 0, \frac{d}{f}\right), \quad P_3 = \left(\frac{a}{b}, 0, 0\right), \quad P_4 = \left(\frac{af - cd}{bf + ce}, 0, \frac{ae + bd}{bf + ce}\right).$$

The first three points are always feasible, P_4 is feasible for

$$af > cd. \qquad (2)$$

Then there is coexistence $P_5 = (R_5, U_5, F_5)$. Its population values are obtained solving for F and U respectively the second and third equations in (1), thus giving

$$F_5 = \frac{1}{k}\left[\lambda(R_5 - s) - \omega\right], \quad U_5 = \frac{1}{h}\left[d + eR_5 - fF_5\right].$$

Substituting into the first one, we obtain the quadratic equation $W(R) \equiv \sum_{k=0}^{2} a_k R^k = 0$ whose roots give the values of R_5. Its coefficients have the following values

$$a_2 = \frac{\lambda}{h}\left(\frac{f}{k}\lambda - e\right) - b - \frac{c}{k}\lambda, \quad a_0 = \frac{1}{hk}\left(dk + fs\lambda + f\omega\right)\left(s\lambda + \omega\right),$$
$$a_1 = a + \frac{c}{k}(s\lambda + \omega) + \frac{1}{hk}[(s\lambda + \omega)(ek - f\lambda) - \lambda(dk + fs\lambda + f\omega)].$$

Now, since $a_0 > 0$, if the parabola $W(R)$ is concave one positive root will exist. Thus a sufficient condition for the existence of P_5 is $a_2 < 0$, i.e., explicitly,

$$f\lambda^2 < h[(b + e)k + c\lambda]. \tag{3}$$

For feasibility, we need also the other population values at a nonnegative level, a fact which is attained for U_5 if $ek > f\lambda$, else we must impose it, giving

$$R_5 < \frac{dk + f\lambda + f\omega}{f\lambda - ek}, \tag{4}$$

as we do for F_5 to obtain

$$R_5 > s + \frac{\omega}{\lambda}. \tag{5}$$

The Jacobian of (1) is

$$J = \left[\begin{array}{ccc} a - 2bR - \lambda U - cF & -\lambda(R - s) + \omega & -cR \\ \lambda U & \lambda(R - s) - kF - \omega & -kU \\ eF & -hF & d + eR - hU - 2fF \end{array}\right]$$

The eigenvalues for P_1 are $-\lambda s - \omega$, d, a, entailing its instability. Those for P_2 are $-(dk + f\lambda s + f\omega)f^{-1}$, $-d$, $(af - cd)f^{-1}$ giving the stability condition

$$af < cd. \tag{6}$$

Comparing this condition with (2), we observe that there is a transcritical bifurcation, for which P_4 emanates from P_2 when the latter becomes unstable. In other words, introducing the healthy prey invasion number

$$R^{(i)} \equiv \frac{af}{cd}. \tag{7}$$

we have that for $R^{(i)} > 1$ the healthy prey establish themselves in the environment.

For P_3 the eigenvalues are $(bd+ae)b^{-1}$, $(\lambda a - \lambda sb - b\omega)b_1$, $-a$, giving instability.

At P_4 one eigenvalue is easily factored out,

$$\frac{\lambda(af-cd)-k(bd+ae)}{ce+bf} - \lambda s - \omega,$$

while the remaining ones are roots of the quadratic equation

$$T(\delta) = \delta^2 + b_1\delta + b_2 = 0, \tag{8}$$

where letting $D = ce + bf$,

$$\begin{aligned} b_1 &= \frac{t_1}{D}, \quad b_2 = \frac{t_3}{D}, \quad t_1 = af(b+e) + bd(f-c), \\ t_3 &= (bd+ae)(af-cd), \quad t_2 = t_1^2 - 4t_3(bf+ce). \end{aligned}$$

Explicitly,

$$T_{1,2} = \frac{-b_1 \pm \sqrt{b_1^2 - 4b_2}}{2} = \frac{-t_1 \pm \sqrt{t_2}}{2(ec+bf)}. \tag{9}$$

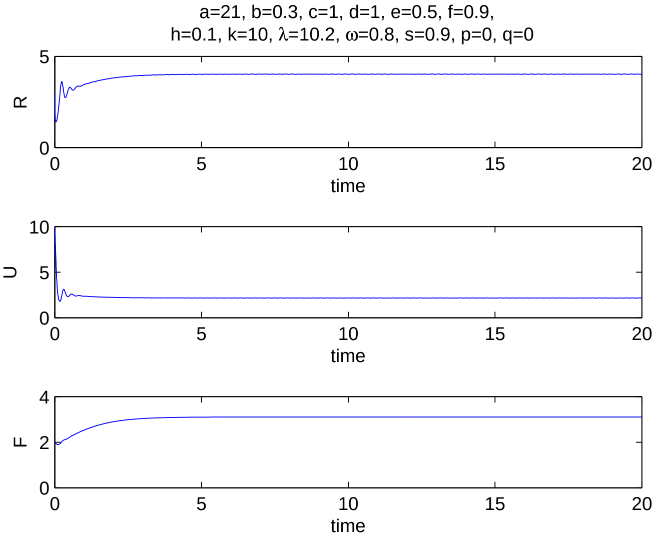


Figure 1: The coexistence equilibrium is attained for the following choice of parameters: $a = 21$, $b = 0.3$, $c = 1$, $d = 1$, $e = 0.5$, $f = 0.9$, $h = 0.1$, $k = 10$, $\lambda = 10.2$, $\omega = 0.8$, $s = 0.9$.

By the feasibility condition (2), $t_3 > 0$ so that $t_2 < t_1^2$. Hence both roots of (9) have negative real part. Stability hinges then just on the first eigenvalue, i.e. $\lambda R_4 < kF_4 + \lambda s + \omega$ or explicitly the following condition

$$\lambda \frac{af - cd}{bf + ce} < k \frac{ae + bd}{bf + ce} + \lambda s + \omega. \tag{10}$$

For the coexistence equilibrium P_5, we have run some simulations to show that it can be attained at a stable level. Figure 1 shows one such instance, for the parameter values $s = 0.9$ and

$$\begin{gathered} a = 21, \quad b = 0.3, \quad c = 1, \quad d = 1, \quad e = 0.5, \quad f = 0.9, \\ h = 0.1, \quad k = 10, \quad \lambda = 10.2, \quad \omega = 0.8. \end{gathered} \tag{11}$$

Here the R_5 equilibrium value is much higher than the number of individuals s that can take cover in the safety niche. Observe also that the same inequality holds also for all the healthy prey population values before attaining the equilibrium level.

3 The case of a cover for the infected

Assume now that part of the infected are somehow confined in an environment in which healthy prey cannot enter. In this way the contagion risk is reduced. Let p denote the fixed number of infected that inhabit the unreacheable territory. With the remaining notation similar to model (1), the system in our present case reads

$$\begin{aligned} \frac{dR}{dt} &= R[a - bR - cF - \lambda(U - p)] + \omega U \\ \frac{dU}{dt} &= \lambda(U - p)R - U[kF + \omega] \\ \frac{dF}{dt} &= F[d + eR - fF - hU] \end{aligned} \tag{12}$$

Again, here we have to remark that for $U < p$ the contributions to the infected class is to be understood to drop to zero. In such case, once again, the infected prey in the system vanish, and the system settles to any equilibrium of the classical disease-free predator-prey model $\widetilde{P}_4 \equiv P_4$, [7].

For (12) the equilibria are again the origin $\widetilde{P}_1 \equiv P_1 = (0, 0, 0)$ and the point $\widetilde{P}_2 \equiv P_2$ but here we find a new predator-free point, while coexistence of healthy prey and predators is forbidden. We thus have

$$\widetilde{P}_0 = \left(\frac{a}{b}, \frac{a\lambda p}{a\lambda - b\omega}, 0 \right).$$

The latter is feasible for $a\lambda > b\omega$, i.e. introducing the disease basic reproduction number R_0, if

$$R_0 \equiv \frac{a\lambda}{b\omega} > 1. \tag{13}$$

The presence of the coexistence equilibrium $\widetilde{P}_5 = (\widetilde{R}_5, \widetilde{U}_5, \widetilde{F}_5)$ can be discussed as follows. From the last equation of (12) we solve for F

$$\widetilde{F}_5 = \frac{1}{f}(d + eR - hU)$$

and substitute into the remaining equations to obtain two conic sections

$$\Phi(R,U) \equiv \frac{k}{f}hU^2 - e\frac{k}{f}RU + U\left(\lambda - \frac{k}{f}d - \omega\right) - p\lambda = 0,$$
$$\Psi(R,U) \equiv -\left(b + \frac{c}{f}e\right)R^2 + \left(\frac{c}{f}h - \lambda\right)RU + \left(p\lambda - \frac{c}{f}d + a\right)R + \omega U = 0,$$

of which we seek an intersection $(\widetilde{R}_5, \widetilde{U}_5)$ in the first quadrant. We study the each one of them separately.

The implicit function $\Phi = 0$ can be solved as a function $R = \rho(U)$,

$$\rho(U) \equiv \frac{1}{fU}\left[khU^2 + (f\lambda - dk - f\omega)U - fp\lambda\right].$$

The numerator is a convex quadratic, which has two real roots with opposite signs, $\widetilde{U}_{\pm}$. In fact Descartes rule shows that independently of the sign of $f\lambda - dk - f\omega$ there is always one variation and one permanence of signs in its coefficients. For $U > 0$ it is therefore a continuous function crossing the U axis at $\widetilde{U}_+ > 0$, that has a vertical asymptote coinciding with the R axis and for $U > 0$ it raises up to infinity, asymptotically approaching the straight line $R = hkf^{-1}U$. Its inverse, $U = \rho^{-1}(R)$, has the R axis as horizontal asymptote for $R \to -\infty$ and goes to infinity for large positive R, crossing the U axis at U_+. This curve corresponds to the level 0 of the surface $\Phi(R,U)$. This implicit function is clearly negative at the origin, since $\Phi(0,0) = -p\lambda < 0$, and by continuity retains this sign everywhere below the curve $U = \rho^{-1}(R)$, while it is positive above it.

The function $\Psi(R,U)$ instead vanishes at the origin. Studying it on the R axis, we find that it must cross it also at the point

$$\widetilde{R}^{(1)} = \frac{fp\lambda - c + af}{bf + ce},$$

which can have either sign. If $\widetilde{R}^{(1)} < 0$, then $\Psi(R,U) > 0$ in the whole first quadrant. Therefore in this case Φ and Ψ do not meet in the first quadrant and the coexistence equilibrium $\widetilde{P}_5$ does not exist.

Conversely, if

$$fp\lambda + af > c \tag{14}$$

since Ψ is a conic section, it must raise up from the origin and then go down to meet the R axis at $\widetilde{R}_1$, thus it defines an arc of a concave function $U \equiv A(R)$ in the first quadrant. We must investigate when this arc $U = A(R)$ and the function $U = \rho^{-1}(R)$ meet. We do so by comparing their respective slopes at $R = 0$. Evidently, since $A(0) = 0$ and $\rho^{-1}(0) = \widetilde{U}_+ > 0$, if $A'(0) < (\rho^{-1})'(0)$, no intersection can exist, recalling the concavity of A and the fact that ρ^{-1} is monotonically increasing, as we can easily verify that $[\rho^{-1}(R)]' = [\rho'(U)]^{-1} > 0$ for $U > 0$. One can also explicitly find the expression of $A(U)$ as

$$A(U) = R\frac{(bf + ce)R + cd - fp\lambda - af}{(ch - f\lambda)R + f\omega}.$$

We must impose the converse condition $A'(0) > (\rho^{-1})'(0)$. Implicit differentiation of Φ and Ψ and evaluation at $R = 0$ yields

$$(\rho^{-1})'(0) = \frac{ekU_+}{2hkU_+ f\lambda - dk - f\omega}, \quad A'(0) = \frac{1}{\omega}(cd - fp\lambda - af).$$

For $U = 0$ we then need to have the slope of A larger than the one of ρ^{-1}, but this does not automatically imply an intersection of the two curves. In fact two intersections exist if we additionally require for instance that at the maximum of the arc A, or in general for any suitable value of the abscissa $\bar{R}$ in $[0, R_1]$, the values of A and ρ^{-1} "interlace", i.e. the following conditions are met

$$A'(0) > (\rho^{-1})'(0), \quad A\left(\bar{R}\right) \geq \rho^{-1}\left(\bar{R}\right). \tag{15}$$

The above conditions are then sufficient for the existence of $\widetilde{P}_5$. In particular we could here easily locate the reference point as $\bar{R} \equiv \frac{1}{2}\widetilde{R}_1$,

The Jacobian of (12) is

$$\widetilde{J} = \begin{bmatrix} a - 2bR - cF - \lambda(U - p) & -\lambda R + \omega & -cR \\ \lambda(U - p) & \lambda R - kF - \omega & -kU \\ eF & -hF & d + eR - hU - 2fF \end{bmatrix}.$$

$\widetilde{P}_1$ is always unstable, since the eigenvalues are d and

$$-\frac{1}{2}\,\omega + \frac{1}{2}\,a + \frac{1}{2}\,\lambda\,p \pm \frac{1}{2}\,\sqrt{\omega^2 + 2\,a\,\omega - 2\,\lambda\,p\,\omega + a^2 + 2\,a\,\lambda\,p + p^2\,\lambda^2}.$$

For $\widetilde{P}_2$ we find the eigenvalue $A_0 = -d$ and

$$\begin{aligned} A_\pm &= \frac{1}{2f}\left[af - kd - \omega f - cd + fp\lambda \pm \sqrt{Y}\right], \\ Y &= 2\,kd\omega\,f + \omega^2 f^2 - 2\,kd^2 c + 2\,\omega\,f^2 a + 2\,kdaf - 2\,\omega\,fcd + 2\,kd\lambda\,pf \\ &\quad -2\,\lambda\,p\omega\,f^2 - k^2 d^2 + 2\,cd\lambda\,pf + 2\,afcd - 2\,af^2\lambda\,p - a^2 f^2 - c^2 d^2 - p^2\lambda^2 f^2. \end{aligned}$$

Stability is then ensured if

$$f(a+p\lambda) < kd + \omega f + cd. \tag{16}$$

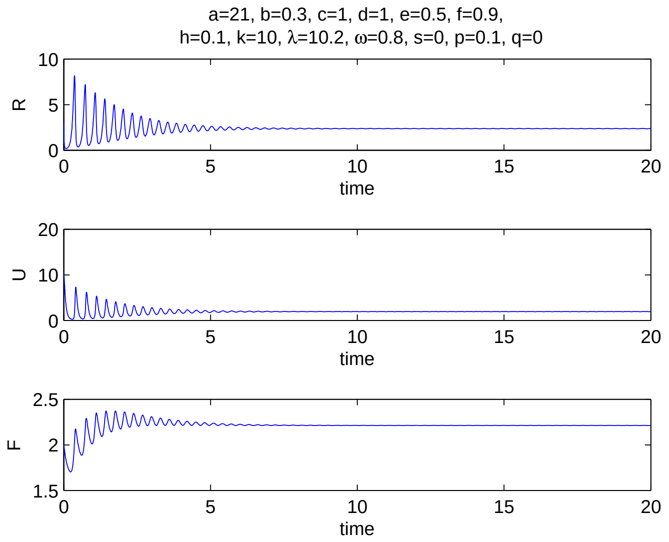


Figure 2: The coexistence equilibrium $\widetilde{P}_5$ is achieved when $p = 0.1$ and the remaining parameters are given by (11) as in Figure 1.

For the point $\widetilde{P}_0$ we have the following eigenvalues

$$B_1 = \frac{dba\lambda - db^2\omega + ea^2\lambda - eab\omega - ha\lambda\, pb}{b\,(a\lambda - b\omega)}$$

and the pair

$$\begin{aligned}
B_\pm &= \frac{1}{2b(a\lambda - b\omega)}\left[a^2\lambda^2 - a\lambda\, b\omega - b\omega\, a\lambda + b^2\omega^2 - ba^2\lambda + ab^2\omega - \lambda\, pb^2\omega \pm \sqrt{X}\right],\\
X &= b^2\omega^2a^2\lambda^2 - 2\,b^3a^3\lambda\,\omega + \lambda^2p^2b^4\omega^2 + a^2b^4\omega^2 + b^2a^4\lambda^2 + b^4\omega^2\omega^2 + 2\,a^4b\lambda^3\\
&\quad + a^2\lambda^2b^2\omega^2 - 2\,a^3\lambda^3b\omega - 2\,a^3\lambda^3b\omega - 2\,ab^4\omega\,\omega^2 - 2\,a^3b^2\omega\,\lambda^2 + 2\,a^2b^3\lambda\,\omega^2 + a^4\lambda^4\\
&\quad + 4\,a^2\lambda^2b^2\omega\,\omega - 2\,a\lambda\,b^3\omega^2\omega - 2\,b^3\omega^2a\lambda\,\omega - 2\,ab^4\omega^2\lambda\,p - 4\,a^3b^2\lambda^2\omega + 4\,a^2b^3\omega\,\lambda\,\omega\\
&\quad - 2\,a^2\lambda^3pb^2\omega + 2\,a\lambda^2b^3\omega^2p - 2\,b^4\omega\,\omega^2\lambda\,p + 2\,\lambda^2pb^3a^2\omega + 2\,b^3\omega\,a\lambda^2p\omega.
\end{aligned}$$

Using feasibility (13), stability in this case is ensured by the following set of conditions

$$a\lambda(db + ea - bhp) > db^2\omega + eab\omega, \quad a\lambda[a\lambda - 2b\omega - ab] < b\omega[bp\lambda - b\omega - ab]. \tag{17}$$

With the help of some simulations we can show that the coexistence equilibrium can be stably achieved, Figure 2. The refuge parameter used is $p = 0.1$ while all the remaining ones are those (11) as in Figure 1. Note that in this case raising the niche level to $p = 0.4$ causes the infected population at some point to fall below this threshold, so that they are wiped out, Figure 3. So while we stated that the disease-free point is not an equilibrium of (12) per se, in suitable situations it would certainly occur. In fact when the infected population U becomes smaller than the level p, and this occurs pretty early in the simulation as observed in Figure 3, the sound prey first and then also the predator populations suddenly surge to finally settle to the coexistence equilibrium of the underlying demographic model.

4 The case of reduced contacts

We consider now another situation, in which we assume that it is the rate of contacts between infected and susceptibles that gets somewhat reduced, due to the effect of a protective niche. In this case then we introduce the fraction $0 \leq q \leq 1$ of avoided contacts. The model, using

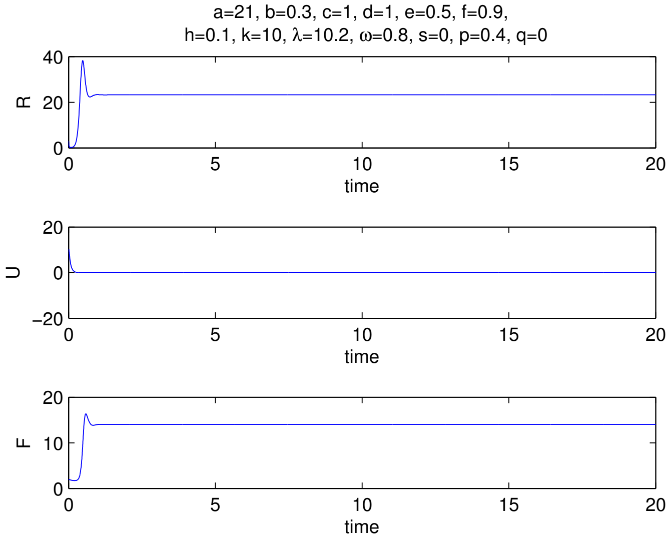


Figure 3: The disease-free equilibrium is attained for $p = 0.4$ with the remaining parameters given by (11) as in Figure 1. Note that the diseased population U falls below the level p very soon, and consequently both the healthy prey first and subsequently the predators pick up, and finally settle to the coexistence equilibrium of the underlying demographic model.

again the very same previous notation, now becomes

$$\begin{aligned} \frac{dR}{dt} &= R[a-bR-cF-(1-q)\lambda U]+\omega U \\ \frac{dU}{dt} &= U[(1-q)\lambda R-kF-\omega] \\ \frac{dF}{dt} &= F[d+eR-fF-hU] \end{aligned} \tag{18}$$

Clearly, by redefining $\beta=(1-q)\lambda$ for $\omega=0$ we get the same model studied in [7]. For the convenience of the reader we summarize the basic results on the equilibria in which at least one of the population vanishes and then extend the study for the coexistence, to encompass here the situation $\omega\neq 0$ not considered in [7] for this specific equilibrium. The equilibria are again all the equilibria of the system (1), namely the origin $\widehat{P}_1\equiv P_1\equiv\widetilde{P}_1$, and $\widehat{P}_2\equiv P_2\equiv\widetilde{P}_2$, $\widehat{P}_3\equiv P_3$, $\widehat{P}_4\equiv P_4$. For feasibility of $\widehat{P}_4$ clearly we need again (2).

Coexistence $\widehat{P}_5=(\widehat{R}_5,\widehat{U}_5,\widehat{F}_5)$ is obtained by solving the second equation in (18) at equilibrium and substituting into the third equation of (18) to get

$$\widehat{F}_5=\frac{(1-q)\lambda\widehat{R}_5-\omega}{k},\quad \widehat{U}_5=\left(\frac{e}{h}-\frac{f}{hk}(1-q)\lambda\right)\widehat{R}_5+\frac{d}{h}+\frac{f}{hk}\omega,$$

and finally from the first equation in (18) we get the quadratic $\sum_{k=0}^{2}c_kR^k$, whose roots determine the value of $\widehat{R}_5$, with $c_0=(dk\omega+f\omega^2)(hk)^{-1}>0$ and

$$c_2=\left(\frac{c}{k}-\frac{e}{h}\right)(1-q)\lambda+\frac{f}{hk}(1-q)^2\lambda^2-b,\quad c_1=a+\frac{c}{k}\omega+\frac{e}{h}\omega-(1-q)\lambda\left(\frac{d}{h}+2\frac{f}{hk}\omega\right).$$

Again we can apply Descartes' rule to have at least a positive root. This occurs for one root if we impose either one of the alternative conditions

$$c_2<0,\quad c_1<0;\qquad c_2<0,\quad c_1>0, \tag{19}$$

and we get two positive roots if

$$c_2>0,\quad c_1<0. \tag{20}$$

We do not write explicitly these conditions. For feasibility we must impose

$$\widehat{R}_5>\frac{\omega}{(1-q)\lambda k} \tag{21}$$

and the condition

$$\widehat{R}_5>\frac{dk+f\omega}{ek-f(1-q)\lambda},\quad ek>f(1-q)\lambda, \tag{22}$$

since the opposite one $ek < f(1-q)\lambda$ would give a negative value for $\widehat{R}_5$.

For $\widehat{P}_1$ the eigenvalues are $-\omega$, d, a, showing its instability.

The eigenvalues of $\widehat{P}_2$ are $-(dk + f\omega)f^{-1}$, $-d$, $(af - cd)f^{-1}$, for which the stability condition is (6). Here again comparing (6) with (2) we observe the existence of a transcritical bifurcation, for which the same conclusions, using the healthy prey invasion number (7) can be drawn as for the model with refuge for the healthy prey (1).

The eigenvalues of $\widehat{P}_3$ are $(bd + ae)b^{-1}$, $[(1-q)\lambda a - b\omega]b^{-1}$, $-a$, thus it is unstable.

For $\widehat{P}_4$ one eigenvalue can easily be factored out, while the other ones are the roots of the quadratic (8). Thus, as found formerly, by feasibility (2) both its roots have negative real part, and stability depends only on the first eigenvalue, namely it is given by $(1-q)\lambda R_4 < kF_4 + \omega$, a condition that can also be explicitly written as

$$(1-q)\lambda\frac{af - cd}{bf + ce} < k\frac{ae + bd}{bf + ce} + \omega. \tag{23}$$

Figure 4 shows the result of a simulation with the same parameter values (11) as for Figure 1, but for $q = 0.1$, assessing the stability of the coexistence equilibrium $\widehat{P}_5$.

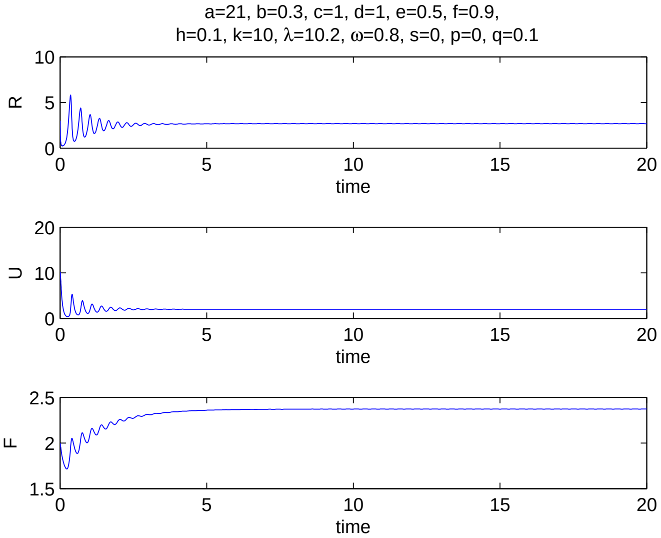


Figure 4: The coexistence equilibrium is attained $\widehat{P}_5$ for the same parameters (11) as in Figure 1 with $q = 0.1$.

5 Models Comparison

The classical predator-prey model underlying these ecoepidemic systems is obtained by eliminating the variable U and its corresponding equation in (18). The resulting system, which can be seen as the projection of the ones considered here onto the disease-free $R - F$ phase plane, has the following equilibria:

$$Q_1 = (0,0), \quad Q_2 = \left(0, \frac{d}{f}\right), \quad Q_3 = \left(\frac{a}{b}, 0\right), \quad Q_4 = \left(\frac{af - cd}{bf + ce}, \frac{ae + bd}{bf + ce}\right).$$

The latter is feasible when (2) holds.

Q_1 and Q_3 are both unstable, in view of their respective eigenvalues a, d and $-a$, $(ae + bd)b^{-1}$. For Q_2 we find $(af - cd)f^{-1}$, $-d$ showing that it is stable exactly when (6) holds. The eigenvalues of Q_4 are complex conjugate, with negative real part, so that Q_4 is unconditionally stable. Being the only such equilibrium, local stability implies global stability. This fact could be shown also via a suitable Lyapunov function.

Thus, the ecoepidemic system exhibits a similar range of behaviors as the demographic ecosystem: coexistence is allowed, both with and without infected, compare P_4 and P_5, and also the predators-only equilibrium P_2, recalling that other food sources for them are available. Evidently, in this prey-free environment, the role of the refuge for the prey is nonexistent. The same does not occur, not surprisingly either, for the disease-free equilibrium P_4. In fact the population levels are not affected by the size of the refuges in any model, but the stability of this equilibrium does in fact depend on this parameter. The way in which the refuges' parameters s and q appear in the stability conditions differs, compare (10) and (23). But both have a stabilizing effect for the ecoepidemic system, a result which as mentioned agrees with former findings in the literature for predator-prey models, [3]. In the case of the reduced contacts model, the refuge favors stability since, mathematically, the left hand side becomes smaller due to a positive q, while in the case of a refuge for the healthy prey it is the right hand side that gets increased by the presence of s. However, since q is a fraction, denoting the relative reduction in the frequency of contacts, while s represents the number of refuges, it is more likely that the latter has a more marked influence on stability.

Note further that the disease-free equilibrium P_4 does not exist per se if the infected find cover, i.e. in system (12). However, we have seen that this equilibrium is achievable when the infected population value falls below the threshold given by the size of the niche p. For the same model (12), however, in place of the disease-free equilibrium, we find an additional situation that does not arise in the other models, in which namely the predators get wiped out from the environment while the prey thrive with their disease becoming endemic. This predator-free environment can be achieved if the conditions (13) and (17) hold. In such situation note that the infected level is directly proportional to the size p of the niche available for their segregation. In particular, if the disease is unrecoverable,

$\omega = 0$, or if there is no intraspecific competition among the healthy prey, $b = 0$, the size of surviving infected is exactly p. If these situations are not met, then the resulting number of thriving infected is larger than p. Hence, the higher the refuges, the more endemic the disease remains, when the predators are wiped out. This is a somewhat counterintuitive result. It is true that the niches help the infected not to get in contact with the susceptibles, but then one would expect also an advantage for the healthy individuals. Instead we find them at the level ab^{-1} which would be attained at the unstable equilibrium P_2. Hence, another way of looking at this situation is to observe that in this case the niche stabilizes the otherwise unstable predator-free equilibrium, at the price of making the disease endemic.

The numerical experiments with the coexistence equilibria of the three models show that using the set of demographic parameter values in (11), i.e. those given by the first row, the system settles to the demographic disease-free equilibrium $(23.2475, 0, 14.0261)$, whose projection onto the $R - F$ phase plane corresponds of course to the equilibrium of the underlying classical predator-prey system, $(23.2475, 14.0261)$. If we now introduce the disease, with the related parameter values found in the second row of (11), we find the ecoepidemic equilibrium $(2.1133, 1.8658, 2.0819)$. As we can easily observe, the disease has a large impact on the system, reducing both its populations by an order of magnitude. Although the epidemics affects only the prey, its effect is felt also by the predators. This can easily be interpreted, because a reduced food supply, due to a lower prey population caused by the disease, must reduce also the predator population and, in addition, consumption of infected prey is harmful for the predators. In other words, diseases, as stated many times in ecoepidemiological research, affect the whole ecosystems, and therefore in environmental studies they cannot be easily neglected.

Coming back to the effects of our safety refuges, we have run simulations using the previous parameter values (11), with various sizes for the refuge coefficients s, p and q. As remarked earlier the proviso holds, that in the models (1) and (12) a check is implemented, for which when $U < p$ and $R < s$ the next to last term in the first equation and the first one in the second equation are set to zero in both (1) and (12). The results are reported in Figures 5-7.

Comparison of the results indicates that for the healthy refuge, the healthy prey and the predators at equilibrium increase in a linear fashion their numbers as s grows, while the infected appear to reach a plateau. When the infected prey have a cover, there is a threshold value of its size p beyond which the disease disappears and the other populations suddenly jump to the level of the corresponding demographic, disease-free, classical model and stay there independently of the value of p. A similar result holds also when it is the contact rate that gets reduced, i.e. for model (18). In this case the equilibria behavior before the threshold value of q is reached appears to be smoother than in the previous case of system (12).

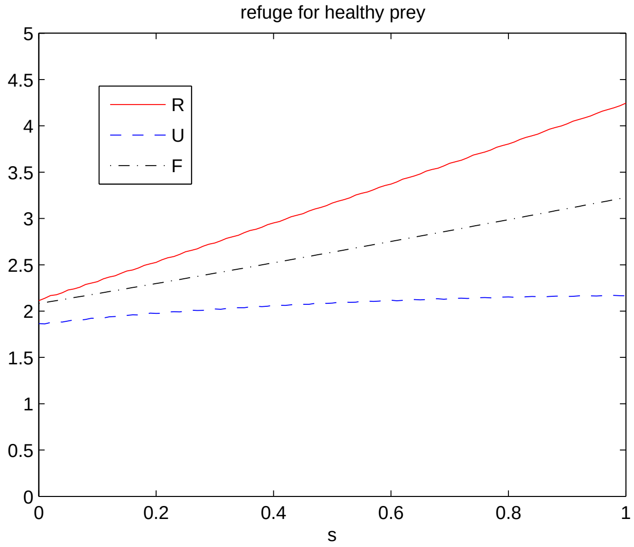


Figure 5: Equilibrium population values of system (1) as function of the refuge size s.

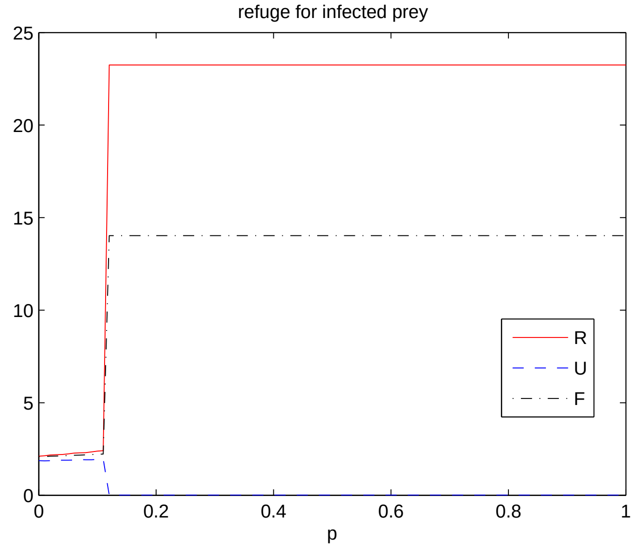


Figure 6: Equilibrium population values of system (12) as function of the refuge size p.

We also plot the equilibrium levels of the various populations as function of the disease parameters λ and ω versus the refuge parameters s, p and q in Figures 8-13

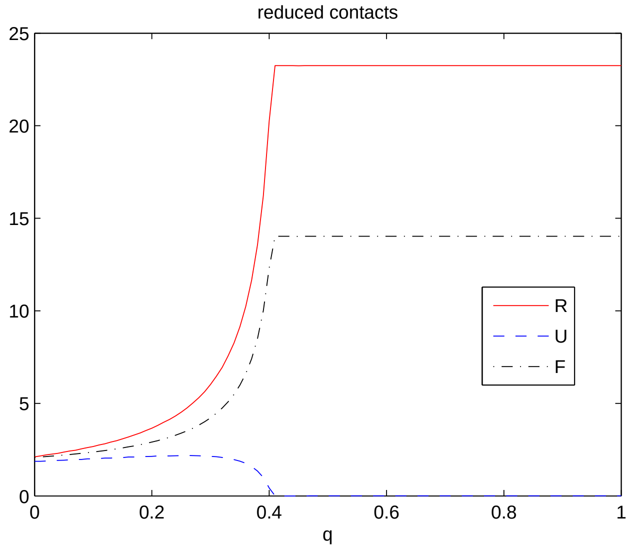


Figure 7: Equilibrium population values of system (18) as function of the contact rate reduction coefficient q.

Comparing the susceptible levels in Figure 8, when the contact rate is high, an improvement in the equilibrium value is obtained for larger value of the refuge s in model (1), while for (12) and (18) an increase in the refuge size is irrelevant, the equilibrium configuration is determined essentiall by the contact rate λ. A similar behavior holds for the predators as well, Figure 10. A corresponding opposite effect is noted among the infected, Figure 9. In (1) a larger s smoooths out the growth of the equilibrium value, which is much sharper for the other two models, once the contact rate crosses the critical threshold.

Considerations along the above lines can be also made when comparing the refuge usage versus the disease recovery rate ω. Comparing Figures 11-13, we see the marked similarities between the equilibrium surfaces of the models (12) and (18), for all the populations involved. Both healthy prey and predators show a linear increase as function of the recovery rate, while the niche apparently does not play any essential role. The infected prey instead seem to reach a plateau. Instead, for the model 1, we find again a linear increas in terms of ω, but what is more important, also a sharp increase of healthy prey and predators as function of the niche size s. A corresponding decrease of infectives can also be observed, which is more marked for high values of the niche size and of the recovery rate, as it should be expected.

Based on these overall considerations, it appears that the model (1) shows the best characteristics in terms of disease reduction. Thus in this type of predator-prey ecoepidemic system with disease just in the prey, for an endemic disease, the ecosystem with a place where some of the healthy individuals can be segregated from coming in contact with disease carriers would exhibit the best features to preserve the epidemics to spread. This result could possibly give some hints to field ecologists as how to fight diseases in wild populations, in case some artificial refuges, unreachable by the diseased individuals, can be provided in specific real-life situations.

References

[1] J. B. Collings, *Bifurcations and stability analysis of a temperature-dependent mite predator-prey interaction model incorporating a prey refuge*, Bulletin of Mathematical Biology **57** (1995) 63-76.

[2] E. González-Olivares, R. Ramos-Jiliberto, *Dynamic consequences of prey refuges in a simple model system: more prey, fewer predators and enhanced stability*, Ecological Modelling **166** (2003) 135-146.

[3] E. González-Olivares, R. Ramos-Jiliberto, *Comments to the effect of prey refuge in a simple predator-prey model*, Ecological Modelling **232** (2012) 158-160.

[4] E. González-Olivares, B. González-Yañez, R. Becerra-Klix, *Prey refuge use as a function of predator-prey encounters*, private communication, submitted to International Journal of Biomathematics (2012).

[5] H. Malchow, S. Petrovskii, E. Venturino, *Spatiotemporal patterns in Ecology and Epidemiology*, CRC, Boca Raton, 2008.

[6] G. D. Ruxton, *Short term refuge use and stability of predator-prey models*, Theoretical Population Biology **47** (1995) 1-17.

[7] E. Venturino, *Epidemics in predator-prey models: disease among the prey*, in O. Arino, D. Axelrod, M. Kimmel, M. Langlais: *Mathematical Population Dynamics: Analysis of Heterogeneity, Vol. one: Theory of Epidemics*, Wuertz Publishing Ltd, Winnipeg, Canada, (1995) 381-393.

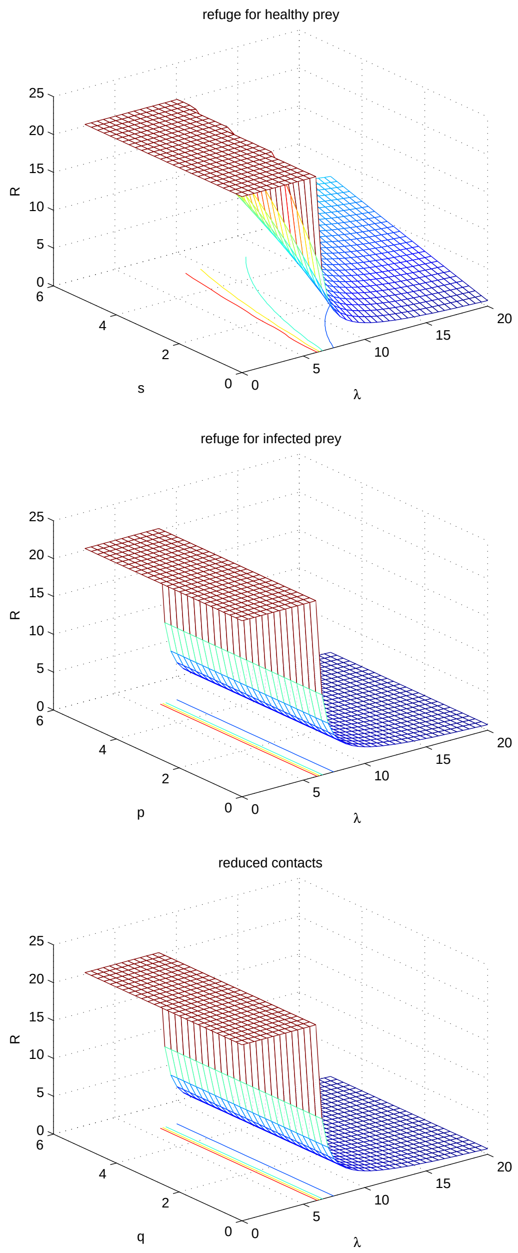


Figure 8: Equilibrium population value for the healthy prey as function of the disease contact rate λ and refuge size: top, refuge size s in model (1); middle, refuge size p in model (12); bottom, reduced contact rate q in model (18). Other parameter values as in (11).

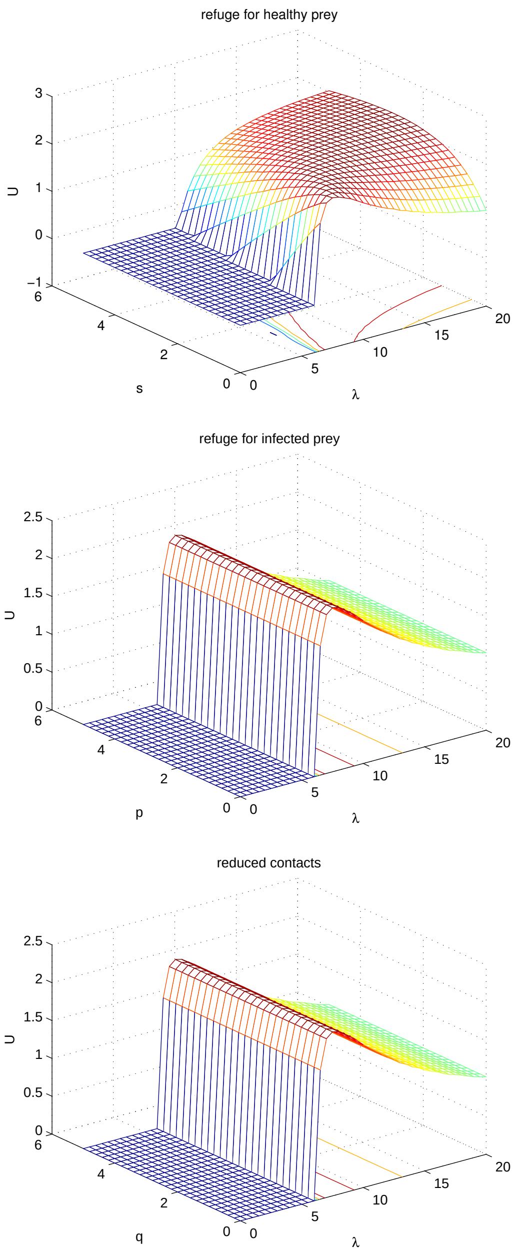


Figure 9: Equilibrium population value for the infected prey as function of the disease contact rate λ and refuge size: top, refuge size s in model (1); middle, refuge size p in model (12); bottom, reduced contact rate q in model (18). Other parameter values as in (11).

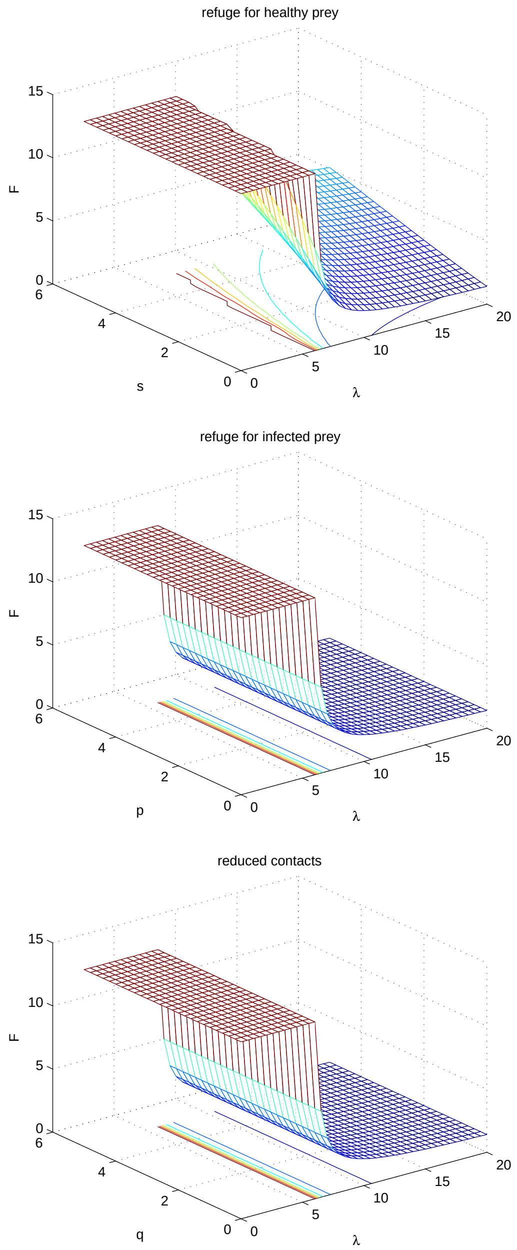


Figure 10: Equilibrium population value for the predators as function of the disease contact rate λ and refuge size: top, refuge size s in model (1); middle, refuge size p in model (12); bottom, reduced contact rate q in model (18). Other parameter values as in (11).

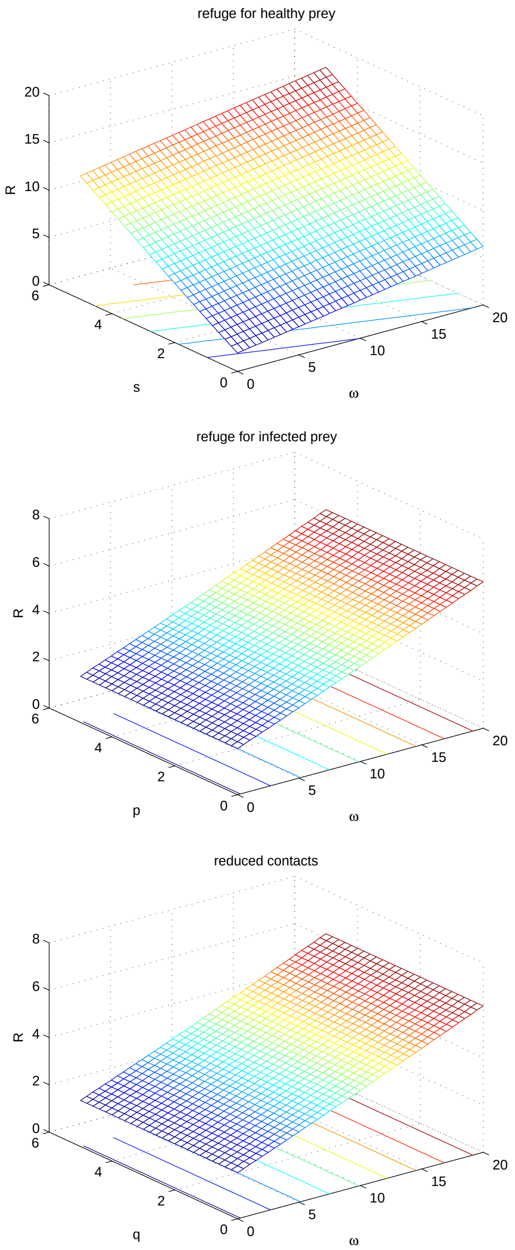


Figure 11: Equilibrium population value for the healthy prey as function of the disease contact rate ω and refuge size: top, refuge size s in model (1); middle, refuge size p in model (12); bottom, reduced contact rate q in model (18). Other parameter values as in (11).

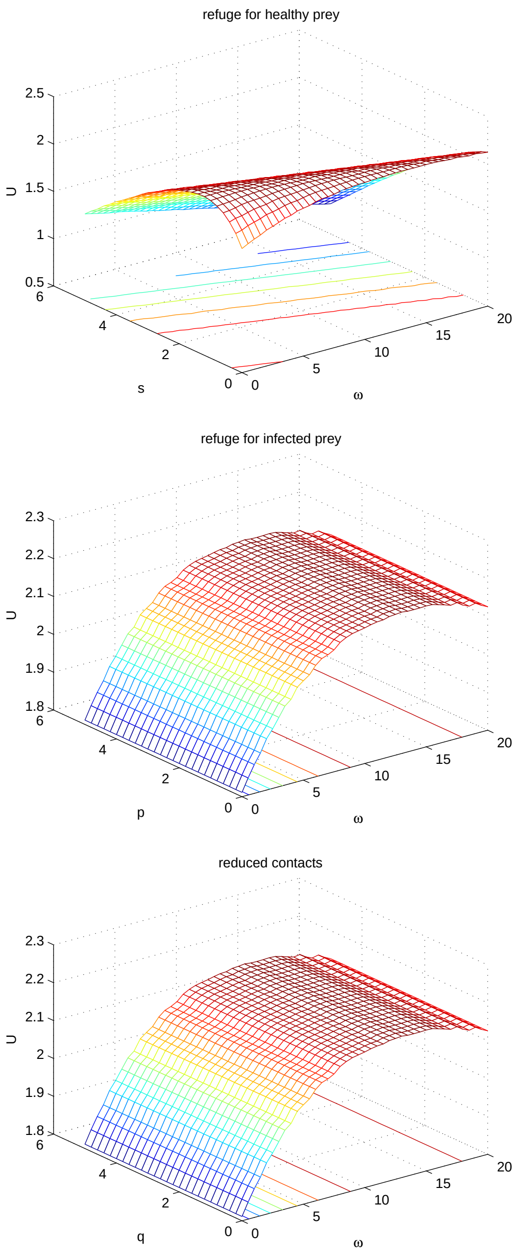


Figure 12: Equilibrium population value for the infected prey as function of the disease contact rate ω and refuge size: top, refuge size s in model (1); middle, refuge size p in model (12); bottom, reduced contact rate q in model (18). Other parameter values as in (11).

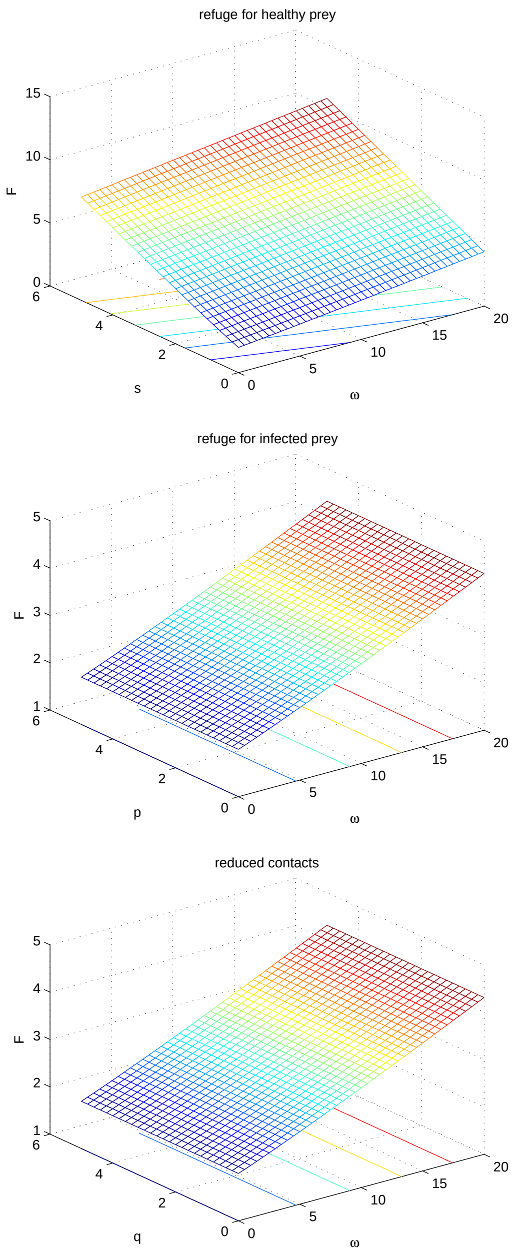


Figure 13: Equilibrium population value for the predators as function of the disease contact rate ω and refuge size: top, refuge size s in model (1); middle, refuge size p in model (12); bottom, reduced contact rate q in model (18). Other parameter values as in (11).

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

GPU Acceleration of a Tool for Wind Power Forecasting

Marcel Burdiat[1], José Ignacio Hagopian[1], Juan Pablo Silva[1], Ernesto Dufrechou[1], Alejandro Gutiérrez[2], Martín Pedemonte[1], Gabriel Cazes[2] and Pablo Ezzatti[1]

[1] *Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay*

[2] *Instituto de Mecánica de los Fluidos e Ingeniería Ambiental, Facultad de Ingeniería, Universidad de la República, Uruguay*

emails: `mburdiat@gmail.com`, `jsign.uy@hotmail.com`, `silvaljp01@hotmail.com`, `edufrechou@fing.edu.uy`, `aguti@fing.edu.uy`, `mpedemon@fing.edu.uy`, `agcm@fing.edu.uy`, `pezzatti@fing.edu.uy`

Abstract

Uruguay is currently undergoing a gradual process of inclusion of wind energy in its matrix of electric power generation. In this context, a computational tool for wind power forecasting has been developed in order to predict the electrical power that will be injected into the electrical grid. The tool is based on the Weather Research and Forecasting (WRF) numerical model, which is the computational bottleneck of the application. For this reason, and in line with several successful efforts of other researchers, this paper presents advances in porting the WRF to GPU. In particular, we present and study the implementation of `sintb` and `bdy_interp1` routines on GPU. The results obtained on a Nvidia's GTX 480 GPU show speedup values of up to 10× when compared with the sequential WRF and almost 5× when compared with the four-threaded WRF. This improvement impacts in a 10% reduction in the total runtime of the WRF.

Key words: Wind power, WRF, GPU, `sintb` *routine,* `bdy_interp1` *routine.*

1 Introduction

Although windmills have been used to generate electricity in a domestic scale, mainly in rural properties, they have not been considered in Uruguay as an alternative energy source until recently. The Wind Power Program (WPP) [3], established in 2007, aims to significantly increase the generation of electricity from wind energy encouraging the installation of large-scale wind farms in different parts of the country. For this reason, several projects are under development, and it is estimated that the wind energy penetration factor will be more than 20% in 2018 [2]. Under the WPP, UTE, the public utility responsible for the generation and transfer of electrical energy, installed the wind farm complex *Ing. Emanuele Cambilargiu.* After three years of continuous operation, it produces an average of 40% of its capacity.

Wind energy has different features compared with other sources of energy.Wind energy is not easy to be stored as potential energy, so the kinetic energy of the air flow is transformed into rotational mechanical energy which is used for electric power generation that has to be injected into the electrical grid right away. Moreover, wind energy has a large variability due to the nature of its origin. For this reason, and due to the importance that wind power will have for electric power generation in Uruguay, it is necessary to have computational tools to forecast the generated power. These tools would allow to take decisions in the management of the electrical grid in order to meet the energy demand.

With this goal, a computational tool for wind power forecasting of the UTE's wind farm complex was developed. The forecasting process runs four times a day and involves the use of the WRF numerical model [8, 16], which is the computational bottleneck of the whole process. The process generates a low resolution forecast for the whole country and a high resolution forecast where the wind farm complex is located, which are publicly available on [4]. This process involves a runtime of approximately two hours on a server with eight cores. This runtime is considered high, mainly because it will become necessary could be desirable to increase the number of wind farms that are predicted.

In recent years, the execution time of several numerical models have been significantly reduced through porting the numerical model computation, either partially or fully, to Graphics Processing Units (GPUs) [6]. Some of the most relevant numerical models that have been successfully ported to GPU can be consulted on [17].

In this paper we study the implementation of `sintb` and `bdy_interp1` routines on GPU in order to accelerate the tool for wind power forecasting. Our experiments show that just porting these routines to GPU is possible to reduce a 10% the total runtime of the WRF. The rest of the paper is structured as follows. In Section 2, we describe the tool for wind power forecasting and the WRF. Then, in Section 3, we review the related works. In Section 4, we introduce the methodology followed for porting routines from CPU to GPU and also describe the implementation of the routines on GPU. In Section 5 we present experimental results and, finally, we discuss some conclusions and future work in Section 6.

2 Tool for Wind Power Forecasting

The forecasting process performed by the computational tool involves several steps. In the first place, the tool obtains public global meteorological forecasts generated by the

NOAA/National Weather Service using the general circulation model of the atmosphere Global Forecast System. These forecasts are generated four times a day, at 00:00, 06:00, 12:00 and 18:00 GMT and have a grid resolution of $1^\circ \times 1^\circ$ (approximately 100km×100km). The WRF is executed locally in our server and uses this data, as well as other parameters as inputs to generate a low resolution forecast for the whole country and one high resolution forecast for the region of the wind farm complex. A forecast of the wind energy fed into the grid by each wind turbine is produced using the predictions of the wind speed and direction at the height of each turbine shaft generated by the WRF, and corrections of these predictions through model output statistics. These results are published automatically in 6-hour intervals [4], having four daily forecasts that predict the outcome for the next 48 hours.

The computationally most expensive step of the whole process corresponds to the WRF execution, which requires approximately two hours on a server with eight cores. Therefore, the performance of the WRF should be tackled in order to increase the performance of the whole tool. A description of the WRF is presented below.

2.1 Weather Research and Forecasting Model (WRF)

The WRF is a numerical model for weather prediction and an atmospheric simulation system for research and operational applications [8]. The WRF model is Eulerian (uses a fixed coordinate system with respect to the earth), non-hydrostatic (includes explicit equations to calculate the pressure and the gravitational force on the vertical axis) and compressible (considers density variations suffered by the various fluids involved). The WRF code is written in Fortran and C, and it currently has about half a million lines of code (version 3.4.1). The WRF execution partitions the domain into rectangular patches that can be assigned to different processors. These patches are subdivided into tiles that can be executed on different threads. Usually, patches need data from other neighboring patches, so in the border tiles it is necessary to propagate changes between different patches.

The tool for wind power forecasting uses the ARW (“The Advanced Research WRF”) core of the WRF, which was primarily designed for research purposes, but it is also used for weather prediction. The WRF generates forecasts of W-E and N-S components that are used to calculate the magnitude of the velocity in m/s and the direction in degrees. To that end, four nested grid levels (30km×30km, 10km×10km, 3.3km×3.3km and 1.1km×1.1km) are generated, where levels of grid with a higher resolution assimilate the information generated by the model in the levels of grid with a lower resolution. The values predicted by the tool have been confirmed with real speed and direction data measured by UTE using anemometers and wind vanes installed in different parts of the country. It is noteworthy that WRF is executed in parallel in eight cores using the shared memory parallelism configuration, which uses the OpenMP API for parallelization.

3 Related work

This section is centered on reviewing previous efforts on porting WRF routines to GPU. The website [10] collects some of the pioneering works on this subject.

In the seminal paper on this subject, Michalakes and Vachharajani [9] address the acceleration of WRF Single Moment 5 Cloud Microphysics (WSM5) module on CUDA C. Although WSM5 only represents 0.4% of the total WRF code, it usually takes up to a quarter of the total processing time on a single core. Each GPU-thread computes the calculations corresponding to a point of the grid, since the calculations of the vertical column of each grid point are independent of other grid points. The authors evaluated their implementation in a domain with a grid resolution of 71×58 points and 27 vertical levels, using a NVidia GeForce 8800 GTX connected to an Intel Pentium-D at 2.8GHz. The numerical results showed slight differences between the GPU and CPU versions, but the visualization outputs of both versions are indistinguishable. The GPU implementation runs 17× faster than the single-threaded CPU version, including the time of the transferences between the host and the device. From this result the authors estimate that it translates in a 1.25× reduction in the total application runtime.

Later, the same authors [11] ported the WRF Fifth Order Positive Definite Tracer Advection. In this case, it is not possible to make a dimensional division of the problem, as in the previous work. Since the number of floating point operations per memory access is only 0.76 and a large number of tracers are usually executed, the GPU implementation overlaps the calculations of a tracer with the asynchronous data transfer required for next tracer. The authors evaluated their implementation in a domain with a grid resolution of 134×110 points, 35 vertical levels and 81 tracers, using a Tesla C1060 GPU connected to a quad-core Intel Xeon E5440. The GPU implementation runs 6.7× faster than the single-threaded CPU version.

Finally, Michalakes et al. [7] evaluated the parallelization of Regional Acid Deposition Model version 2 (RADM2) module on a multicore processor (two quad-core Intel Xeon 5400), a GPU (Tesla C1060) and a Cell Broadband Engine Architecture (CBEA) [5] device (PowerXCell 8i). RADM2 requires to run the Rosenbrock's algorithm, which involves constructing a Jacobian matrix and a LU decomposition, independently for each grid point. In the GPU implementation, some calculations have to be computed on the CPU due to the large amount of memory required, even though the fundamental steps of the algorithm are executed on the GPU. The parallel implementations were evaluated in a domain with a grid resolution of 40×40 points and 20 vertical levels. The speedup with respect to a single-threaded CPU version was 7.5× for the multicore processor, 8.5× for the GPU and 41× for the CBEA. Despite failing to achieve high speedup values for the GPU implementation, the authors note that the GPU is the cheapest platform of the three used and they also highlight the simplicity of its programming compared to CBEA.

In another line of work, Ruetsch et al. [18] ported the Long-Wave Rapid Radiative

Transfer Model (RRTM) module to CUDA Fortran. As in other works, the implementation exploits the independence between the different vertical columns of the domain, even achieving a higher level of independence in some cases, resulting in a finer grain parallelism. The authors evaluated their implementation in a domain with a grid resolution of 73×60 points and 27 vertical levels, using a Tesla C1060 GPU connected to a quad-core Intel Xeon E5440 at 2.83 GHz. The speedup of the GPU implementation versus a single-threaded CPU version is between 8× and 10×. The numerical results showed similar results between the GPU and CPU versions.

Huang et al. developed other relevant works on the subject, in which several modules were ported to CUDA C, with the final objective of completely porting WRF to GPU. In one of their works [12], the authors port the Goddard Shortwave Radiation module exploiting the independence between the different vertical columns of the domain, as the other authors. The GPU implementation also uses two streams in order to overlap kernel computation on the GPU with transferences between the host and the device. The authors evaluated their implementation in a domain with a grid resolution of 433×308 points and 35 vertical levels, using two GeForce GTX 590, i.e. four GPUs, connected to a six-core Intel i7 970. The speedup is 116× for the GPU implementation versus a single-threaded CPU version, including the time of the transferences between the host and the device.

Following the same idea, the authors ported the Stony Brook University 5-Class Cloud Microphysics module to GPU [13]. In this work, the authors used single precision floating point arithmetic and compiled using the CUDA flag `-use_fast_math` that uses faster but less accurate computation. Using the same scenario and the same platforms as in the previous work, the speedup value reported is 352×. Huang et al. also worked on porting the WSM5 [14], WDM5 Cloud Microphysics [19], Purdue Lin Cloud Microphysics [20] and Kessler Cloud Microphysics [15] modules to GPU using the same ideas, scenario and platforms as in the previous works, reporting speedup values of 357×, 147×, 156× and 70× respectively.

While speedups reported by Huang et al. are impressive, there are some aspects that should be taken into account in order to explain these values. In the first place, the scenario used by Huang et al. is more than 30 times larger than the ones used by other authors, which undoubtedly contributes to the improvement in the speedup. Another aspect that certainly impacts on these speedup values reported is the use of the CUDA compile flag `-use_fast_math`. Finally, an important issue is that the criterion followed by other authors to select which routines port to GPU is based on the importance of the routine for the execution of the WRF, while it is not clear the criterion used by Huang et al.

4 Improving the Performance of the WRF

Our work is focused on improving the runtime of the WRF, and in particular the modules required by the forecasting tool developed for the wind farm complex *Ing. Emanuele Cam-*

bilargiu. Therefore, an empirical approach is used to determine in which module concentrate our effort. To that end, we follow the guidelines suggested by Michalakes and Vachharajani [9], and Delgado et al. [1] to port applications developed in Fortran to CUDA C.

Four different steps are identified to port an application from Fortran to CUDA C: profiling, development, testing and optimization. The profiling stage aims to determine which are the modules or routines that require longer execution times. Then, in the development stage, it is recommended to port first Fortan to C, and then C to CUDA, since directly porting Fortran to CUDA can incorporate many errors due to the different characteristics of the languages involved. In the testing stage, it has to be evaluated that the results obtained by the GPU are similar to the ones obtained by the CPU, taking into account that there might be floating point rounding differences caused by the parallelism. Finally, in the optimization stage, the CUDA implementation is fine tuned in order to reduce the runtime.

The tools for profiling applications follow two completely different philosophies. On the one hand, there are tools that make changes in the source code at compile time, including extra code to obtain metrics of the application execution. On the other hand, there are tools that work like a debugger or a virtual machine that obtain the metrics in run-time using the original code. Since both approaches are complementary, we decided to profile the WRF using one tool of a kind. We use *gprof*, which follows the former approach and it is considered the de-facto profiler in academic environments since it is included in the GNU project, and *valgrind* that follows the latter approach. The reports of the single-threaded execution of WRF generated with both tools showed a great consistency in the results, despite following different philosophies. The routine with longer runtime was `sintb` that required around the 15% of the total execution time. For this reason, we focus our work in porting this routine to GPU. Even though each invocation of `sintb` does not take a large runtime, the routine is invoked more than 10 million times during the WRF execution.

4.1 sintb and bdy_interp1 Routines

To begin with, we study the feasibility of porting `sintb` routine to GPU analyzing the relationship between transference and computation time for a call to `sintb`. We found that this ratio was four to one, so we examined `bdy_interp1` routine, which is the only routine that calls `sintb`. `bdy_interp1` makes four consecutive calls to `sintb` with the same input matrices, but with different sets of positions. For this reason, a single transference is required to make the four invocations to `sintb` thus compensating the high relative computational cost of the transference. So, it was considered a better option to also port code from the `bdy_interp1` routine to the GPU instead of only porting `sintb`.

After the four `sintb` calls, the `bdy_interp1` routine makes assignments to other matrices using the results from the invocations. Both invocations and assignments are in a loop parallelized using OpenMP in the original code, but the set of positions of each invocation to `sintb` does not depend on the iteration step. Therefore, we decided to group all calls to `sintb` in a single call, postponing all assignments for the end. Following that approach, the Fortran code of `sintb` was ported to C, and it was verified that the results were similar to those obtained by the original version. Then, the C code of `sintb` was ported to CUDA.

Now, the main bottlenecks were the transference of the result matrices (25% of the total runtime) and the final assignments to other matrices in `bdy_interp1` (30% of the total runtime), since they could not be run concurrently with the kernel. For this reason, we chose to migrate the final assignments in `bdy_interp1` to a second kernel in CUDA, not only to reduce the computation time of this task, but also taking benefit from the fact that all the inputs of the routine are already on the GPU. This brought a couple of other improvements. On the one hand, the final results of `bdy_interp1` are nearly half of the size of the intermediate results produced by the invocation to `sintb`, so the transference time from GPU to CPU is significantly reduced. On the other hand, the matrices where the final results are updated can be copied from the GPU to the CPU asynchronously and concurrently with the execution of the first kernel, which hides this transference time. The pseudocode of the host side of the CUDA implementation outlined above is presented in Algorithm 1.

Algorithm 1 `bdy_interp1` Host Side Pseudocode

```
synchronous transferences required by sintb, host to device, stream 1
asynchronous transferences to be updated with final results, host to device, stream 2
invoke sintb kernel through stream 1
synchronize stream 2                 ▷ Blocks host until all calls in stream 2 are completed
invoke second kernel through stream 1                        ▷ Updates the final results
synchronous transferences of final results, device to host, stream 1
```

The memory accesses of the GPU implementation are coalesced and the memory spaces allocated on the GPU are reused in order to reduce calls to malloc and free. Finally, we performed an analysis to find the best balance between the number of registers and shared memory used per block. The best configuration consists in using only 24KB of shared memory per block, allowing to execute up to two concurrent blocks on each multiprocessor.

5 Experimental Results

This section describes first the scenario used for the experimental study and the execution platform. Then, the results obtained are presented and analyzed.

5.1 Test Instance

The test instance consists in real information of the 07/15/2012 for a 12 hours climate forecast of the wind farm complex *Ing. Emanuele Cambilargiu.* The test instance is considered representative of a normal day of operation of the park. The numerical model uses four nested grid levels of 30km×30km, 10km×10km, 3.3km×3.3km and 1.1km×1.1km that assimilate information from the lower resolution to the higher resolution levels of the grid. The 30km×30km domain has a grid resolution of 74×61 points and 54 vertical levels, while the others domain has a grid resolution of 40×40 points and 54 vertical levels.

5.2 Test Environment

The execution platform for the CPU runs is a PC with a dual core Intel Core i3-2100 processor at 3.10 Ghz with 8 GB using the Fedora 15 Linux operating system. It should be noted that the processor supports hyper-threading. The CPU versions were compiled using the `-O3` flag. The execution platform for the GPU runs is an Nvidia's GeForce GTX 480 (480 CUDA cores, Fermi architecture, Compute Capability 2.0) connected to the CPU platform. The GPU versions were compiled using the Nvidia's CUDA compiler 4.1 version with the `-O3` flag. All the reported total runtime of the GPU executions always include the transference time of data between CPU and GPU.

5.3 Experimental Analysis

We begin our analysis with the parallelization of the WRF using OpenMP. Table 1 presents the experimental results regarding the performance obtained. The table includes the execution time of the single-threaded, two-threaded and four-threaded (using hyper-threading) WRF, as well as the speedup values of the parallel executions. The speedup value obtained using two threads is not close to linear; this shows that the parallelization of the WRF is not a trivial task and that an adequate load balance has to be achieved in order to benefit from additional processing units. The execution time with two and four threads allows us to affirm that in this case the use of hyper-threading is justified.

Table 1: OpenMP Parallelization of WRF

Single Thread	Two Threads		Four Threads	
Runtime (mins)	Runtime (mins)	Speedup	Runtime (mins)	Speedup
226.70	126.77	1.79	103.38	2.19

Table 2 presents the execution time of the single-threaded and four-threaded (using hyper-threading) of `bdy_interp1` routine, as well as the speedup value of the parallel version. The runtimes reported are the average of all the executions of the routine (the routine is executed more than 2.5 million times). The speedup value obtained with four threads is slightly worse than for the entire application which may indicate that the routine has further difficulty to scale with a multithreaded parallelization.

We continue our analysis with the parallelization of the WRF using a GPU. In the first place it should be noted that there are no significant differences regarding the numerical

Table 2: OpenMP Parallelization of `bdy_interp1`

Single Thread	Four Threads	
Runtime (ms)	Runtime (ms)	Speedup
24.00	11.50	2.09

accuracy between the results obtained by the GPU and the CPU implementation. There are several alternatives to evaluate the performance of the GPU implementation. The two key aspects that influence the definition of speedup that can be considered are how many threads run on the CPU implementation taken as a reference and whether the transference time is included in the runtime of the GPU implementation or not. Four different definitions of speedup arise from the combination of both aspects: parallel CPU run and including the transference time (Eq. 1), parallel CPU run and not including the transference time (Eq. 2), single-threaded CPU run and including the transference time (Eq. 3) and single-threaded CPU run and not including the transference time (Eq. 4).

$$\text{Speedup}_{I} = \frac{\text{Runtime of 4-threaded } \texttt{bdy\_interp1}}{\text{GPU total runtime of } \texttt{bdy\_interp1}} \quad (1) \qquad \text{Speedup}_{II} = \frac{\text{Runtime of 4-threaded } \texttt{bdy\_interp1}}{\text{GPU total runtime of } \texttt{bdy\_interp1}} \quad (2)$$

$$\text{Speedup}_{III} = \frac{\text{Runtime of 1-threaded } \texttt{bdy\_interp1}}{\text{GPU total runtime of } \texttt{bdy\_interp1}} \quad (3) \qquad \text{Speedup}_{IV} = \frac{\text{Runtime of 1-threaded } \texttt{bdy\_interp1}}{\text{GPU total runtime of } \texttt{bdy\_interp1}} \quad (4)$$

All the previous works reviewed in the related work section use Speedup$_{III}$ to evaluate the performance of the GPU implementation. Huang et al. [12, 13, 14, 15, 19, 20] also report Speedup$_{IV}$ since they claim that when the WRF is fully ported to GPU it would not be necessary to transfer the data. We believe that the speedup values, which are calculated against the four-threaded CPU execution (Speedup$_{I}$ and Speedup$_{II}$), are fairer to evaluate the performance of the GPU implementation because it is a more realistic execution scenario since it is a little artificial to limit the computing capacity of the CPU for the evaluation. Anyway, we include Speedup$_{III}$ and Speedup$_{IV}$ in our analysis in order to be able to compare our results with other works.

Table 3 presents the runtime of the GPU implementation of the `bdy_interp1` routine with the transference and computation time disaggregated, as well as the four speedup values described above. To begin with, the transference time represents 20% of the total execution time so the increase in the speedup when the transference is not considered is only 20%. Although it is a small percentage of the total time, it can be explored the alternative of using asynchronous transferences instead of synchronous ones, in order to overlap them with computation on the GPU or the CPU, thus reducing the time involved in the transferences. In the second place, the Speedup$_{I}$ value obtained when compared with the four-threaded execution of the WRF is 4.89. An almost five times reduction in the runtime of the routine can be considered a good result, since the improvement is accomplished versus an already parallel implementation that was developed by the staff of developers of the WRF. Finally, the Speedup$_{III}$ value obtained is 10.21 that is comparable to the values reported by Michalakes et al. and Ruetsch et al. We could not reach similar values to the ones reported by Huang et al., but the scenario we used is significantly smaller and fast math was not used due to the characteristics of the routine that we ported to GPU.

Table 3: GPU Parallelization of `bdy_interp1`

Runtime (ms)			Speedups			
Transference	Computation	Total	Speedup$_I$	Speedup$_{II}$	Speedup$_{III}$	Speedup$_{IV}$
0.45	1.90	2.35	4.89	6.05	10.21	12.63

Table 4 presents the runtime of the single-threaded and four-threaded WRF with `bdy_interp1` routine executing on the GPU, as well as the corresponding speedup values. The implementation of this routine in GPU produces a reduction of more than 10 minutes (10.26% of the total runtime) in the runtime of the four-threaded WRF and of 20 minutes (8.75% of the total runtime) in the runtime of the single-threaded WRF.

Table 4: GPU Parallelization of WRF

Single Thread + `bdy_interp1` on GPU		Four Threads + `bdy_interp1` on GPU	
Runtime (mins)	Speedup	Runtime (mins)	Speedup
206.87	1.10	92.77	1.11

6 Conclusions and Future Work

In this paper we have studied the acceleration of a tool for wind power forecasting on a GPU. To this end, we profiled the WRF model, which is the main bottleneck of the application, and determined that `sintb` is the routine with longer runtime (around 15% of the total runtime of WRF). A further analysis showed that it was a better alternative besides porting `sintb` to GPU also porting `bdy_interp1` routine. The implementation of `sintb` and `bdy_interp1` routines on Nvidia's GTX 480 GPU obtained speedup values of up to 10× and almost 5× when compared with the single-threaded and four-threaded execution on CPU, respectively. The acceleration obtained implies a 10% reduction in the total runtime of the WRF.

We identify three lines for future work. The first one is to identify additional routines to port to GPU in order to accelerate the tool for wind power forecasting. This leads to a second line of interest, which consists in executing the WRF with the two routines on the GPU for a whole forecast for the next 48 hours. Finally, we aim to use as execution platform a GPU with Kepler architecture that recently came to market.

Acknowledgements

G. Cazes, E. Dufrechou, P. Ezzatti, A. Gutiérrez and M. Pedemonte acknowledge support from Agencia Nacional de Investigación e Innovación (ANII), Uruguay and Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE), Uruguay. E. Dufrechou, P. Ezzatti and M. Pedemonte also acknowledge support from Programa de Desarrollo de las Ciencias Básicas (PEDECIBA), Uruguay.

References

[1] J. Delgado, G. Gazolla, E. Clua and S. Sadjadi, *A Case Study on Porting Scientific Applications to GPU/CUDA*, Journal of Computational Interdisciplinary Sciences **2**, (2011) 3–11.

[2] *Dirección Nacional de Energía - Ministerio de Industria, Energía y Minería (in Spanish)*, `http://www.dne.gub.uy/`.

[3] Dirección Nacional de Energía - Ministerio de Industria, Energía y Minería, *Programa de Energía Eólica en Uruguay (in Spanish)*, `http://www.energiaeolica.gub.uy/`.

[4] A. Gutiérrez and G. Cazes, *Pronósticos Numéricos Operativos en Uruguay (in Spanish)*, `http://www.fing.edu.uy/cluster/eolica/`.

[5] C. Johns and D. Brokenshire, *Introduction to the Cell Broadband Engine Architecture*, IBM Journal of Research and Development **51(5)**, (2007) 503–519.

[6] D. Kirk and W. Hwu, *Programming Massively Parallel Processors, Second Edition: A Hands-on Approach*, Morgan Kaufmann, 2012.

[7] J. Linford, J. Michalakes, M. Vachharajani and A. Sandu, *Multi-core Acceleration of Chemical Kinetics for Simulation and Prediction*, SC '09, Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, (2009) 1–11.

[8] J. Michalakes, J. Dudhia, D. Gill, T. Henderson, J. Klemp, W. Skamarock and W. Wang, *The Weather Research and Forecast Model: Software architecture and performance*, 11th ECMWF Workshop on HPC in Meteorology, (2004) 156–168.

[9] J. Michalakes and M. Vachharajani, *GPU Acceleration of Numerical Weather Prediction*, Parallel Processing Letters **18(4)**, (2008) 531–548.

[10] J. Michalakes and M. Vachharajani, *GPU Acceleration of Numerical Weather Prediction*, `http://www.mmm.ucar.edu/wrf/WG2/GPU/`.

[11] J. Michalakes and M. Vachharajani, *GPU Acceleration of Scalar Advection*, Available online: `http://www.mmm.ucar.edu/wrf/WG2/GPU/Scalar_Advect.htm`.

[12] J. Mielikainen, B. Huang, H.A. Huang and M.D. Goldberg, *GPU Acceleration of the Updated Goddard Shortwave Radiation Scheme in the Weather Research and Forecasting (WRF) Model*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **5(2)**, (2012) 555–562.

[13] J. Mielikainen, B. Huang, H.A. Huang and M.D. Goldberg, *GPU Implementation of Stony Brook University 5-Class Cloud Microphysics Scheme in the WRF*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **5(2)**, (2012) 625–633.

[14] J. Mielikainen, B. Huang, H.A. Huang and M.D. Goldberg, *Improved GPU/CUDA Based Parallel Weather and Research Forecast (WRF) Single Moment 5-Class (WSM5) Cloud Microphysics*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **5(4)**, (2012) 1256–1265.

[15] J. Mielikainen, B. Huang, H.A. Huang and M.D. Goldberg, *Compute Unified Device architecture (CUDA)-based Parallelization of WRF Kessler Cloud Microphysics Scheme*, Computers & Geosciences **52**, (2013) 292–299.

[16] National Center for Atmospheric Research, *Weather Research and Forecasting Model*, `http://www.wrf-model.org/index.php`.

[17] Nvidia, *CUDA Community Showcase*, `http://www.nvidia.com/object/cuda_showcase_html.html`.

[18] G. Ruetsch, E. Phillips and M. Fatica, *GPU acceleration of the Long-wave Rapid Radiative Transfer Model in WRF Using CUDA Fortran*, Available online: `http://www.pgroup.com/lit/articles/nvidia_paper_rrtm.pdf`.

[19] J. Wang, B. Huang, H.A. Huang and M.D. Goldberg, *Parallel Computation of the Weather Research and Forecast (WRF) WDM5 Cloud Microphysics on a Many-Core GPU*, IEEE 17th International Conference on Parallel and Distributed Systems (ICPADS), (2011) 1032–1037.

[20] J. Wang, B. Huang, H.A. Huang and M.D. Goldberg, *GPU Acceleration of the WRF Purdue Lin Cloud Microphysics Scheme*, Proc. SPIE 8183, High-Performance Computing in Remote Sensing, (2011) 81830R.

*Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.*

On recovery of state parameters of systems via video-images analysis

Alexander P. Buslaev[1] **and Marina V. Yashina**[2]

[1] *Department of Higher Mathematics, Moscow Automobile and Road State Technical University*

[2] *Department of Mathematical Cybernetics and Information Technologies, Moscow Technical University of Communications and Informatics*

emails: `apal2006@yandex.ru`, `yash-marina@yandex.ru`

Abstract

Problem of development of pattern recognition methods at the processing of video sequences for the creation of real-time monitoring for complex socio-technical systems is discussed. Qualitative properties of the assumed dynamical systems are strongly dependent on the input data. At the same time, acquisition means of real values of inputs are the most difficult and requires of costly measuring systems.

Therefore, the search for effective and system solutions for processing video sequences in order to restore the characteristics of dynamic systems that are models of complex socio-technical systems are very important.

Key words: Social-technical systems, pattern recognition

1 Introduction

The problem of intellectual monitoring design becomes relevant for research and optimization of functioning of *complex social-technical systems* (STS).

STS modeling is connected with taking into account *a lot of factors* are categorized into two classes: basic and secondary.

The parameters, describing a basic unit, are deterministic component of the model. Other parameters are averaged or become stochastic. Thence in general, models of social and technical systems are mixed, i.e., being *deterministic – stochastic models*, for example, [1].

2 Traffic as STS

Obviously all transport systems of the cities (megalopolises) and regions are examples of complex social – technical systems. Operation of such systems is connected with providing of society safety, then ensure of the equipment, as a rule not cheap equipment, then solution of the ecological problem, etc.

So, in Russia with the 140-million population there are annually about 35 000 killed persons and about 200 000 traumatized persons with various degree weights in traffic accidents.

As result of non sufficient street-road network, traffic jams are increased and cause huge economic losses.

3 Visualization of the dynamical systems

Forecasting of a condition of street traffic in megalopolises becomes a global problem because saturated traffic flows possess instability, because local events influence quickly extends on all network.

One of approaches to modeling consists in video - images processing of traffic flows in megalopolis that been received "from above", i.e., from space, aircraft or dirigible balloons. Large number of information, for example, for Moscow from 200000 to 500000 thousands cars it is necessary to recognize, to process and make recommendations for traffic control.

This problem can not be solved in real time without automation, [2], [3]. And aposteriory processing is not effective. Now "Yandex-probki" system and similar web-services of other known companies give us possibility to look the map of traffic flow in some megalopolises of Russian Federation by means of selective information on a high-speed mode of separate cars. Also even less stable information (as a rule, text mode) about the reasons of movement difficulties, [4]. According to Russian proverb "A picture is worth a thousand words ...", it is represented that information visualization of even the general plan would increase quality of the forecast and regulations very significantly.

4 Street-road network recognition

In the present review of the general plan we will assume that unit of processing is the black and white image of the megalopolis or its part that is essentially not local. Let us consider that capture of information happens in a way of rather high point to consider that the image has constant depth, and not too remote that objects of recognition, in our case they are cars, occupied a quantity of pixels. Standard procedures of smoothing shot are applied in scales correlated to part of the average car (in pixels). As basic element of sampling of the screen we chose an rectangle, that is similar to permission in pixels in a half of the average size of the car,assuming that the main part of components of flow has no more than two colors.

Automatic recognition of traffic flow is carried out by means of the analysis of a variation of function of difference module of average means of gray intensity on time for everyone splitting units. Thus the cells, that containing a movement, [5], are detected $SRN = \{a_{ij}\}$, $(i,\ j) \in IJ$.

5 Estimates of flow numerical characteristics

5.1 Recovery of flow intensity

Function of gray intensity for each cells of the selected set of $SRNof$ will allow to determine intensity of movement as number of transitions of $q_{ij}of$ from "it is free = 0" to "by time is occupied = 1" in unit of time T, $q_{ij} = q_{ij}\ (T)/T$. Summation of q_{ij} on any subset of IJ, forming section, it gives a flow intensity, gives classical characteristic, [4].

5.2 Recovery of flow density

Thus, in each timepoint each cage of IJ is in a condition "0" or "1". For any fixed subset of IJ a share "1" estimates *flow density.* This rough characteristic is applicable to all conditions of movement. However in case of so-called "working movement" it is possible to specify density assessment, counting not quantity of units ("1"), and quantity of clusters from units "1".

5.3 Recovery of the turbulent areas

We fix some cell from IJ. As the network is connected, the quantity of neighboring cells is not empty. Let us calculate correlation functions of states of a considered cell with the next by which it is possible to determine "the movement direction" and secondary flow migrations. Relative size of these sizes characterizes *coefficient of mixed of cars* , a factor which negatively influences on flow work.

5.4 Detection of velocity and movement mode

After definition of movement direction the movement velocities of a busy particle is estimated by means of time, a particle necessary for movement in following cell.

Synchronous movement of a connected chain of cells means that one rigid object is processed. The so-called connected flow means that the strong correlation between schedules of the next cells (clusters) exits. The free mode of movement is accompanied by low correlation between the next objects.

6 Kerner three-phase theory and automated processing of the trajectories field

Traffic flow theory received a a significant stimulus into development, caused widespread of devices for automatically register of motion characteristics.

Now some sections of highway have been equipped such devices, and therefore a large amount of factual information has accumulated.

Kerner [6], [7], has deeply processed of available data, and, on the one hand, demonstrated the inadequacy of existing theoretical approaches, and also formulated some fundamental assumptions about the structure of traffic flows.

In the presented theory a significant place is the visual analysis of *trajectories field* at coordinates (t, x) and $(t, x,\ dotx)$ for individual vehicles.

For example, a connected chain at steady state in a short time Δt gives the following approximate image, Fig.1.

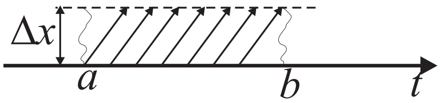


Figure 1: Trajectories fragment of steady state chain

Assuming displacement vectors are the same and small, we obtain that the percentage of pixels occupied by trajectories, is inversely proportional to slope, i.e., velocity. Thus, the image processing of trajectories field allows to separate the areas of movement with given speed mode, including congestion, Fig. 2.

7 Cluster and stochastic flow models on regular networks – chainmails and intelligent monitoring for testing and verification of software

The present method of screen processing by detectors field is relevant for checking the adequacy of the software flow for complex networks.

The present method of screen processing by detectors field is relevant for checking the adequacy of the software flow for complex networks.

One of study problems is to identify the qualitative properties of the flows on the chainmails, i.e. networks of rings, [8].

In particular, the model is composed of a planar ring with shape as shown in Fig. 3. Points N, W, S and E are the nodes adjacent to the ring junctions.

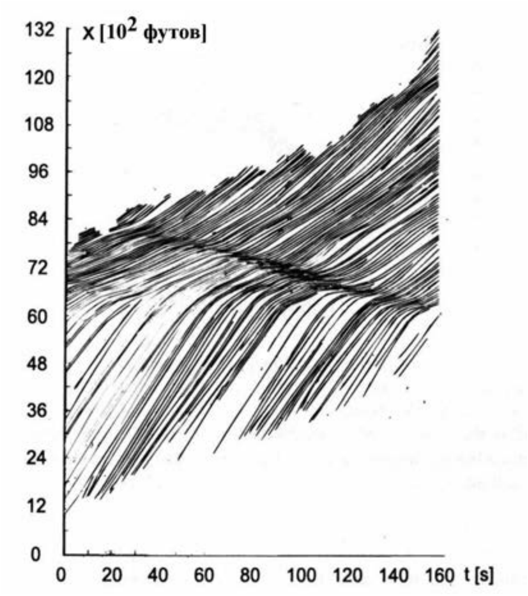


Figure 2: trajectories field, [6]

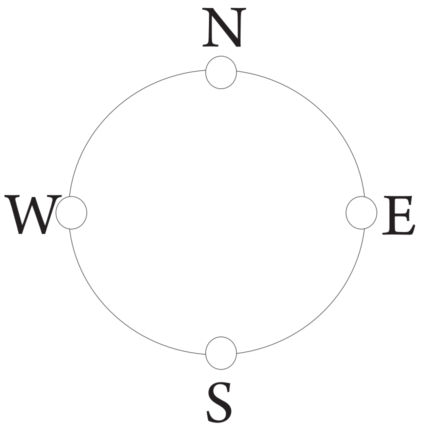


Figure 3: Ring of chainmail

Depending on flow model, either cluster approach or random walks, [9], [10], the flow components move on the ring. They need to interact and migrate to neighboring rings in accordance with the rules strictly.

The software simulates the system status on computer display and intelligent processing allows to find the accordance of the visible system behavior to the given rules.

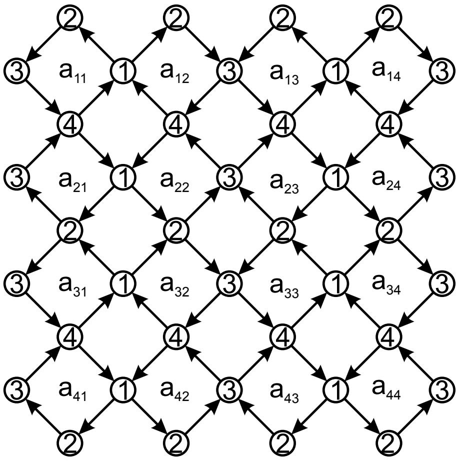

2 2 2 2
3 a_{11} 1 a_{12} 3 a_{13} 1 a_{14} 3
4 4 4 4
3 a_{21} 1 a_{22} 3 a_{23} 1 a_{24} 3
2 2 2 2
3 a_{31} 1 a_{32} 3 a_{33} 1 a_{34} 3
4 4 4 4
3 a_{41} 1 a_{42} 3 a_{43} 1 a_{44} 3
2 2 2 2

Figure 4: Chainmail

In addition, once the credibility of the software is confirmed, it is very useful is the dentify the behaviortrends of a complex dynamical system depending on its parameters.

In particular, for the simplest problem of traffic on chain mail, [11] when on each ring in a certain direction the particle moves only (green), waiting (red), competes for a free cell (yellow), it is relevant of identification of stationary waiting areas (red) after a end-time or non-zero-ergodic area measure.

Acknowledgements

This work has been supported by grants of RFBR No. 12-01-00974-a and No. 11-07-00622-a.

References

[1] BUSLAEV A.P., PRIKHODKO V.M., TATASHEV A.G., YASHINA M.V. *The deterministic-stochastic flow model.* (2005) http://arXiv.org/auth/v.php N 0504139v1

[2] V. KASTRINAKI, M. ZERVAKIS, K. KALAITZAKIS. *A survey of video processing techniques for traffic applications.* Elsevier. J. Image and Vision Computing 21 (2003) 359 – 381.

[3] E. ATKOYCIUNAS, R. BLAKE, A. JUOZAPAVIYCIUS, M. KAZIMIANEC. *Image proccesing in Road Traffic Analysis.* Nonlinear Analysis: Modelling and Control, 2005, Vol. 10, No. 4, 315 – 332.

[4] Bugaev A.S., Buslaev A.P., Yashina M.V. *Road traffic in megalopolis: problems and perspectives of solution. Part 1. General solutions.* - M. Tekhnopoligrafcentr., 2009. – 184.

[5] Buslaev A.P., Provorov A.V., Yashina M.V. *Mathematical recognition problems of particle flow characteristics by video sequence images.* IPCV'13, The 2013 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP'09). Las Vegas, Nevada, USA (July 22-25, 2013), in Proc. of the 2013 Int.Conf. on Image processing, computer vision and pattern recognition (has been accepted)

[6] Kerner B.S. *The Phisics of Traffic. Empirical Freeway Pattern Features, Engineering Applications and Theory.* Springer-Verlag Berlin Heidelberg, 2004, – 682.

[7] Kerner B.S. *Introduction to Modern Traffic Flow Theory and Control*, Springer-Verlag, 2009, – 265.

[8] Kozlov V.V., Buslaev A.P., Tatashev A.G. *Monotonic random walks and clusters flows on networks. Models and applications.* Lambert Academic Publishing, 2013, – 300.

[9] Buslaev A.P., Tatashev A.G. *On exact values of monotonic random walks characteristics on lattices.* Journal of Concrete and Applicable Mathematics (JCAM) Vol. 11, No.1, 2013. 17–22.

[10] Buslaev A.P., Tatashev A.G., Yashina M.V. *Cluster flow models and properties of appropriate dynamic systems.* Journal of Applied Functional Analysis (JAFA), Vol. 8, No. 1, 2013. 54–76.

[11] Buslaev A.P., Tatashev A.G., Yaroshenko A.M., Yashina M.V. *Qualitative Properties of Dynamical System on Toroidal Chainmail* Proc. of the Int. Conf. NAAM13 Rhodes Island, Greece (2013)

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

A predator-prey model with strong Allee effect on prey and interference among predators

Javier Cabrera-Villegas[1], **Fernando Córdova-Lepe**[2] **and Eduardo González-Olivares**[3]

[1] *Departamento de Matemática, Universidad Metropolitana de Ciencias de la Educación*

[2] *Departamento de Matemática, Física y Estadística, Universidad Católica del Maule*

[3] *Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso*

emails: `biomathjcv@gmail.com`, `fcordova@ucm.cl`, `ejgonzal@ucv.cl`

Abstract

In this work a Gause type predator-prey model (as extension of Volterra model) is analyzed. It is assumed that a simple Allee effect affects the prey population and there exists interference among predators. The stability analysis is giving establishing conditions for the existence and local stability of equilibrium points. Results for the extintion and persistence of the species are given.

Key words: Predator-prey model, Allee effect, functional response, stability, interference among predators.

1 Introduction

In this work, a Gause type predator-prey model [8] is analyzed, modifying the well known Volterra model [14], in which the functional response is linear; we introduce two new factors assuming the prey growth rate is affected by a strong Allee effect [7] and there exists competition between predators [1].

The Allee Effect is a positive relationship between any component of individual fitness and either numbers or density of conspecific [6, 12, 13]. This phenomenon has been also called *depensation* in Fisheries Sciences, or *negative competition effect* (*positive density dependence*) in Population Dynamics [5, 10].

Populations can exhibit Allee effect dynamics due to a wide range of biological phenomena, such as reduced antipredator vigilance, social thermoregulation, genetic drift, mating difficulty, reduced antipredator defense, and deficient feeding at low densities; however, several other causes may lead to this phenomenon (see Table 1 in [2] or Table 2.1 in [7]).

The Allee effect can be divided into two main types called *strong Allee effect* [15] or *critical depensation* [5, 10] and *weak Allee effect* [13] or *noncritical depensation* [5, 10].

The strong Allee effect implies the existence of a threshold population level m [1, 4], below which the population becomes extinct. This requires the population growth rate must to be negative for population sizes minor that m.

Many continuous time equations have been used to model the Allee effect [4], although most of them are topologically equivalent [9], *i.e.*, solutions have the same qualitative behavior.

On the other hand, different behavior of predator influencing the interrelation among prey and predators can be assumed, as example: (i) behavior typical of territorial animals where individuals “waste time” in direct contests thereby decreasing time each could otherwise devote to foraging (searching for or handling prey); (ii) spatially aggregated predators and prey and so on [3].

In particular, we consider interference or competition between predators as element affecting the dynamics of the proposed model. *Predator interference* or simply *interference* is a collective term that embraces a number of specific mechanisms [3].

In this work, we model the interference of predator as presented in the book by Bazykin [1] considering the linear functional response independent of predator density, which means that any single predator affects the prey population growth rate independently of its conspecifics. However, to express the action of another predators in the interaction, in our formulation is considered a self-interference term as in the logistic growth rate.

The model and general settings are presented in Section 2. In Section 3, the local nature of equilibrium points and the possibility of: a) extinction, b) extinction of predators only, or c) the possibility of coexistence, according to regions of the parameter space, are presented.

2 The model

Using the simplest way to describe the Allee effect and considering competition (interference) between predators, the model to be studied is described by the planar system:

$$X_\mu : \left\{ \begin{array}{lcl} \frac{dx}{dt} & = & r\left(1 - \frac{x}{K}\right)(x-m)\,x - qxy, \\ \frac{dy}{dt} & = & (px - c - ey)\,y, \;\; (x,y) \in \Omega, \end{array} \right. \tag{1}$$

where $x(t)$ and $y(t)$ represent the prey and predator populations size, respectively at instant $t \geq 0$, the parameters are all positive, *i.e.*, $\mu = (r, K, q, p, c, e, m) \in \mathbb{R}_+^6 \times]0, K[$ and $\Omega = \left\{(x,y) \in \mathbb{R}^2 / 0 \leq x, 0 \leq y\right\}$.

The parameters have the standard ecological meanings. In the first equation of (1): “r” is the intrinsic prey growth rate, “K” is the prey enviromental carrying capacity, “m”

is the strong Allee effect threshold and "q" is the quantity of prey that can be killed by predators in each time unit. In the second one: "p" is the efficiency with which predators convert consumed prey into new predators, "c" is the natural death rate of predators and "e" denotes the struggle in predators [11].

In order to simplify the calculations were carried out a re-parameterization and time scaling. It is proved that (1) is a topologically equivalent to the system that follows:

$$Y_\eta : \left\{ \begin{array}{rcl} \frac{du}{d\tau} &=& [(1-u)(u-M)-v]u, \\ \frac{dv}{d\tau} &=& BE[\frac{u-C}{E}-v]v, \ \ (u,v)\in\bar{\Omega}, \end{array} \right. \tag{2}$$

where $M = \frac{m}{K}$, $B = \frac{p}{r}$, $C = \frac{c}{pK}$ and $E = \frac{er}{pq}$. The new parameter vector is $\eta = (M,B,C,E) \in]0,1[\times\mathbb{R}^3_+$. Moreover, $\bar{\Omega} = \left\{(u,v)\in\mathbb{R}^2/0\le u, 0\le v\right\}$.

3 Results

For any vector of parameters $\mu \in \mathbb{R}^6_+\times]0,K[$ there exist the equilibria: $(0,0)$, $(M,0)$ and $(1,0)$. In addition, some positive equilibrium $(\bar{u},\bar{v})$ at the intersection of the curves of zero per capita growth $(1-u)(u-M)-v=0$ and $u-C-Ev=0$ may appear. Then, the coordinate $\bar{u}$ has to satisfies the quadratic equation $Eu^2+Au+L=0$, with $A=1-ME-E$ and $L=ME-C$. Hence, we have at most two positive equilibria.

By linearization we obtain the following table with the local behavior of the non positive equilibria:

	$(0,0)$	$(M,0)$	$(1,0)$
$M<1<C$	stable	saddle-node	stable
$M<C<1$	stable	saddle-node	saddle-node
$C<M<1$	stable	unstable	saddle-node

(3)

None, one or two positive equilibria exist according to the conditions that follow:

N^o	Cond. 1	Cond. 2	Cond. 3
0	$M<1<C$		
1	$M<C<1$		
0	$C<M<1$	$M>1-E^{-1}$	
0	$C<M<1$	$M<1-E^{-1}$	$A^2<4EL$
2	$C<M<1$	$M<1-E^{-1}$	$A^2>4EL$

(4)

When $M<1<C$ or $M<C<1$, we can consider the stable manifold Γ of $(M,0)$, which is a solution that (making the time to back) goes towards the inside of the first quadrant of the phase plane. Let us denote Γ_+ (resp. Γ_-) the subset of $\mathbb{R}_+\times\mathbb{R}_+$ above and to the left (resp. above and to the rigth) of the curve Γ.

Theorem 1 (Total extintion) *Given $p \in \bar{\Omega}$, the omega-limit set $\omega(p)$ is $\{(0,0)\}$, if some of the conditions that follow is satified:*

- $C \geq 1$ *and* $p \in \Gamma_+$.
- $C < 1$, $A \geq 0$ *and* $L \geq 0$.
- $A^2 - 4EL < 0$.
- *If* $L > \frac{BE-A-1}{2}\bar{u} - \frac{BCE}{2}$, *when* $A < 0$ *or* $L < 0$ *and* $A^2 - 4EL \geq 0$.
- $p \in \Gamma_+$ *and, in addition,* $L < \frac{BE-A-1}{2}\bar{u} - \frac{BCE}{2}$ *with* $C < 1$, $A < 0$ *or* $L < 0$, *and* $A^2 - 4EL \geq 0$.

Theorem 2 (Predator extintion) *A solution $(u(\cdot), v(\cdot))$ satisfies $v(t) \to 0$ as $t \to \infty$, but $\limsup(u) > 0$, if only if $C \geq 1$ and $(u(\cdot), v(\cdot))$ is in Γ_- at some instant.*

Theorem 3 (Coexistence) *A solution $(u(\cdot), v(\cdot))$, $(u(0), v(0)) = p \in \bar{\Omega}$, is such that $\limsup(u), \limsup(v) > 0$ if some of the conditions that follow is satified:*

- $p \in \Gamma_-$ *and moreover* $L < \frac{BE-A-1}{2}\bar{u} - \frac{BCE}{2}$ *with* $C < 1$, $A < 0$ *or* $L < 0$, *and* $A^2 - 4EL \geq 0$.
- *If* $L > \frac{-A}{2}\bar{u}$ *when* $A < 0$ *or* $L < 0$ *and* $A^2 - 4EL \geq 0$ *and p is over the the stable manifold of the positive equilibrium point.*

Acknowledgment. This work was partially financed by the Fondecyt 1120218 and DIEA-PUCV 124.730/2012 projects.

References

[1] A. D. BAZYKIN, *Nonlinear dynamics of interacting populations*, Nonlinear Sciences Series A Vol. 11 World Scientific Publishing Co. Pte. Ltd.1998.

[2] L. BEREC, E. ANGULO AND F. COURCHAMP, *Multiple Allee effects and population management*, Trends in Ecology and Evolution, **22** (2007) 185–191.

[3] L. BEREC, *Impacts of foraging facilitation among predators on predator-prey dynamics*, Bulletin of Mathematical Biology, **72** (2010) 94–121.

[4] D. S. BOUKAL AND L. BEREC, *Single-species models and the Allee effect: Extinction boundaries, sex ratios and mate encounters*, Journal of Theoretical Biology, **218** (2002) 375–394.

[5] C. W. Clark, *Mathematical Bioeconomics. The optimal management of renewable resources*, 2nd edition, John Wiley and Sons, 1990.

[6] F. Courchamp, T. Clutton-Brock and B. Grenfell, *Inverse dependence and the Allee effect*, Trends in Ecology and Evolution, **14** (1999) 405–410.

[7] F. Courchamp, L. Berec and J. Gascoigne, *Allee effects in Ecology and Conservation*, Oxford University Press, 2008.

[8] H. I. Freedman, *Deterministic Mathematical Models in Population Ecology*, Marcel Dekker, 1980.

[9] E. González-Olivares, B. González-Yañez, J. Mena-Lorca and R. Ramos-Jiliberto, *Modelling the Allee effect: are the different mathematical forms proposed equivalents?*, In R. Mondaini (Ed.) Proceedings of International Symposium on Mathematical and Computational Biology, E-papers Serviços Editoriais Ltda. 53–71, 2007.

[10] M. Liermann and R. Hilborn, *Depensation: evidence, models and implications*, Fish and Fisheries, **2** (2001) 33–58.

[11] X.-x. Qiu and H.-b. Xiao, *Qualitative analysis of Holling type II predator-prey systems with prey refuges and predator restricts*, Nonlinear Analysis: Real World Applications, **14** (2013) 1896–1906.

[12] P. A.Stephens and W. J. Sutherland, *Consequences of the Allee effect for behaviour, ecology and conservation*, Trends in Ecology and Evolution, **14** (1999) 401–405.

[13] P. A. Stephens, W. J. Sutherland and R. P. Freckleton, What is the Allee effect?, Oikos, **87** (1999) 185–190.

[14] P. Turchin, *Complex population dynamics. A theoretical/empirical synthesis*, Monographs in Population Biology 35, Princeton University Press, 2003.

[15] G. A. K. van Voorn, L. Hemerik, M. P. Boer and B. W. Kooi, *Heteroclinic orbits indicate overexploitation in predator–prey systems with a strong Allee*, Mathematical Biosciences **209** (2007) 451–469.

Proceedings of the 13th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2013
24–27 June, 2013.

Homogenization of the Poisson equation with Dirichlet conditions in random perforated domains

Carmen Calvo-Jurado[1], **Juan Casado-Díaz**[2] **and Manuel Luna-Laynez**[2]

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

[2] *Department of Differential Equations and Numerical Analysis, University of Sevilla*

emails: `ccalvo@unex.es`, `jcasadod@us.es`, `mllaynez@us.es`

Abstract

Key words: homogenization, random perforated domains
MSC 2000: 35R60, 35B27

1 Extended abstract

For a bounded open set $O \subset \mathbb{R}^N$, and a sequence of open sets $O_\varepsilon \subset O$, we consider the Poisson equation with Dirichlet conditions on ∂O_ε. The problem is to study the asymptotic behavior of the solutions, when ε tends to zero, by finding the equation satisfied by its limit. This allows us to describe the macroscopic behavior of the material corresponding to the mixture of the solid part O_ε and the holes $K_\varepsilon = O \setminus O_\varepsilon$. The obtention of this limit equation is a classical problem which has been considered since the early 80's. Usually, the set K_ε is a union of very small connected components distributed in O. It is well known that the homogenized equation is an elliptic problem in the whole of O which contains in general a new term of order zero depending on the distribution and size of the holes, the Cioranescu-Murat "strange term" ([6]). The most classical result corresponds to the homogenization of the Poisson equation with holes of size $\varepsilon^{\frac{N}{N-2}}$ if $N \geq 3$, or $\varepsilon^{-\frac{1}{\varepsilon^2}}$ if $N = 2$, which are periodically distributed with period ε. In this case the coefficient corresponding to the new term of zero order is a positive constant. The case of arbitrary holes and nonlinear equations has been considered in several papers such as [4], [5], [6], [7], [9], [10], [16].

Our purpose in the present work is to consider the case where the holes are randomly distributed. To simplify, we assume $N \geq 3$, and similarly to the classical periodic setting, we consider holes of size $\varepsilon^{\frac{N}{N-2}}$ such that the distance between two holes is of order ε. Namely:

We consider a probability space $(\Omega, \mathcal{F}, P)$, a subset $\tilde{\Omega} \subset \Omega$, and a function $T(x) : \Omega \to \Omega$, for every $x \in \mathbb{R}^N$, which defines an ergodic measure preserving dynamical system in $\mathbb{R}^N$. Then, for a compact set $K \subset \mathbb{R}^N$, we define the sequence of random holes as

$$K_\varepsilon(\omega) = \bigcup_{z \in \mathbb{R}^N,\, T(z)\omega \in \tilde{\Omega}} \left(\varepsilon z + \varepsilon^{\frac{N}{N-2}} K \right), \quad P\text{-a.e. } \omega \in \Omega.$$

Denoting by $O_\varepsilon(\omega)$ the open set obtained from O by removing the random holes, $K_\varepsilon(\omega)$, we want to study the asymptotic behavior of the solutions u_ε of

$$\left\{ \begin{array}{l} -\Delta_x u_\varepsilon(\omega, x) = f_\varepsilon(\omega, x) \;\text{ in } O_\varepsilon(\omega) \\ u_\varepsilon(\omega, x) = 0 \;\text{ on } \partial O_\varepsilon(\omega) \end{array} \right. \qquad P\text{-a.e. } \omega \in \Omega, \tag{0.1}$$

where f_ε converges strongly in $L^2_P(\Omega; H^{-1}(O))$ to a function f.

The homogenization of random problems has been considered in several papers, see e.g. [2], [3], [8], [11], [12], [14], [15]. In particular, the G. Neguetseng and G. Allaire two-scale convergence method ([1], [13]), which is a very useful tool in periodic homogenization, has been extended in [2] to the setting of random homogenization problems. In general, the heterogeneities considered in those papers are given by functions of the form $F(T(\frac{x}{\varepsilon})\omega)$, with ω taking values on the probability space, T a dynamical system as above and F a random variable. However, in our problem, the homogenization process contains two different sizes ($\varepsilon^{\frac{N}{N-2}}$ and ε) to describe it, and then, it can not be analyzed by the results of [2]. To solve this difficulty, we introduce in the present paper a new extension of the two-scale method, which is based on some ideas used in [4] for the homogenization of Dirichlet elliptic problems in (deterministic) periodic domains. We show that the solution u_ε of (0.1) converges weakly in $L^2_P(\Omega; H^1_0(O))$ to the unique solution u of the problem

$$\left\{ \begin{array}{l} -\Delta_x u(\omega, x)) + \gamma\kappa u(\omega, x) = f(\omega, x) \;\text{ in } O \\ u(\omega, x) = 0 \;\text{ on } \partial O \end{array} \right. \qquad P\text{-a.e. } \omega \in \Omega,$$

where the new term $\gamma\kappa u$ is the equivalent for our random problem of the Cioranescu-Murat strange term in the deterministic case ([6]). Similarly to the classical result, it is given by the capacity κ of the closed set K in $\mathbb{R}^N$, multiplied by γ, the mean density of holes in O. Thanks to the ergodic theory we prove that γ does not depend on $\omega \in \Omega$ or $x \in O$. From the physical point of view this means that the limit behavior of the material corresponding to the mixture of the solid part $O_\varepsilon(\omega)$ and the holes $K_\varepsilon(\omega)$ is deterministic. Similar results but assuming different assumptions about the random distribution of the holes and using different techniques have been obtained in [3] and [14].

Acknowledgements

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] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math Anal. **23** (1992) 1482–151.

[2] A. Bourgeat, A. Mikelic, S. Wright, *Stochastic two-scale convergence in the mean and applications* J. Reine Angew. Math. **456** (1994) 19-51.

[3] L.A. Caffarelli, A. Mellet, *Random homogenization of an obstacle problem*, Ann. Inst. H. Poincaré Anal. Non Linèaire **26** (2009) 2, 375-395.

[4] J. Casado-Díaz, *Two-Scale convergence for nonlinear Dirichlet problems in perforated domains,* Proceedings of the Royal Society of Edinburgh A, **130** (2000) 249–276.

[5] J. Casado-Díaz, A. Garroni, *Asymptotic behavior of nonlinear elliptic systems on varying domains,* SIAM J. Math. Anal. **31** (2000) 581–624.

[6] D. Cionarescu, F. Murat, *Un terme étrange venu d'ailleurs,* in Nonlinear Partial Differential Equations and Their Applications, Collége de France Seminar, Vols. **II** and **III**, Research Notes in Math. 60 and 70 Pitman, London, 1982, 98–138 and 154–178.

[7] G. dal Maso, A. Defranceschi, *Limits of nonlinear Dirichlet problems in varying domains,* Manuscr. Math. **61** (1988) 251–278.

[8] G. Dal Maso, L. Modica, *Nonlinear stochastic homogenization and ergodic theory*, J. Reine Angew, Math. **368** (1986) 28–42.

[9] G. Dal Maso, U. Mosco, *Wiener-criterion and* Γ*-convergence,* Appl. Math. Optim. **15** (1987) 15–63.

[10] G. Dal Maso, F. Murat, *Asymptotic behaviour and correctors for the Dirichlet problems in perforated domains with homogeneous monotone operators,* Ann. Sc. Norm. Sup. Pisa **4** (1997) 239–290.

[11] V.V. Jikov, S.M. Kozlov, O.A. Oleinik *Homogenization of differential operators and integral functionals*, Springer-Verlag, Berlin, 1994.

[12] S.M. Kozlov, *Homogenization of random operators,* Math. U.S.S.R. Sb.**37** (1980) 167–180.

[13] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization,* SIAM J. Math. Anal. **20** (1989) 608–623.

[14] G.C. Papanicolaou, S.R.S. Varadhan, *Diffusion in regions with many small holes,* in *Stochastic differential systems*, Lecture Notes in Control and Information Sci. Springer, Berlin-New York, **25** (1980).

[15] G.C. Papanicolaou, S.R.S. Varadhan, *Boundary value problems with rapidly oscillating random coefficients,* Colloq. Math. Soc. J. Bolyai, North-Holland, Amsterdam-New York, **27** (1981) 835-873.

[16] I.V. Skrypnik, *Averaging of nonlinear Dirichlet problems in punctured domains of general structure,* Mat. Sb. 187, **8** (1996) 125–157.