Proceedings of the 2008 International Conference on Computational and Mathematical Methods in Science and Engineering

Hotel Meliá Galúa
La Manga, Murcia, Spain
13-17, June, 2008

CMMSE
Computational and Mathematical
Methods in Science and Engineering

Editor:
J. Vigo-Aguiar

Associate Editors:
Harihar Khanal(USA), J. L. Garcia Guirao (Spain), Shinnnosuke Oharu (Japan)
and Ezio Venturino (Italy)

ISBN 978-84-612-1982-7
Register Number: 08/8003

Printed on acid-free paper

Volume I

Proceedings of the 2008 International Conference on Computational and Mathematical Methods in Science and Engineering

J. Vigo Aguiar Editor

Preface

We are honoured to bring you this collection of articles and extended abstracts from the *Eighth International Conference on Computational and Mathematical Methods in Science and Engineering* (CMMSE 2008), held at the Hotel Galua, La Manga del Mar Menor, Spain June 13--17, 2008. The primary focus of CMMSE is on new ideas and interdisciplinary interaction in rapidly growing fields of computational mathematics, mathematical modelling, and applications.

CMMSE 2008 special sessions represent advances in numerical solution of ordinary differential equations, mathematical modelling in artificial intelligence, industrial mathematics, algorithms and computations for complex network, bio-mathematics, dynamical systems, computational partial differential equations, computational quantum chemistry, and education reform and use of new teaching resources in mathematics.

In addition to the use of numerical analysis and differential equations techniques to modelling scientific and engineering will also touch other areas where the use of Numerical mathematics is limited, but not the use of Computational Mathematics are, such fields of reasoning or mathematical theory of artificial intelligence are also included in this volume.

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

La Manga del Mar Menor, Spain, June 02, 2008

J. Vigo-Aguiar, Harihar Khanal, J. L. Garcia Guirao, S. Oharu, Ezio Venturino

CMMSE 2008 Special Sessions:

Session Title	Organizers
Numerical Solution of O.D.E.	Guido Vander Berghe & J. Vigo-Aguiar
Mathematical models in artificial intelligence	Juan M. Corchado & Javier Bajo
Industrial Mathematics	Bruce A. Wade
Algorithms and Computation for Complex Networks	Stefano Boccaletti, & Regino Criado
Bio-Mathematics	Ezio Venturino
Computational P.D.E.	Harihar Khanal
Computational Quantum Chemistry	Erkki Brändas
Mathematics, the Education Reform and the use of New Teaching Resources	Regino Criado & J. Vigo Aguiar
Dynamical Systems	Juan L. Garcia Guirao

Acknowledgements:

Many thank to Edward Twizell, Editor-in-Chief of International Journal of Computer Mathematics, and Paul Mezey, Editor-in-Chief Journal of Mathematical Chemistry, who agrees to publish selected full papers of CMMSE-2008. The selection will be based on the international peer review of at least two independent reviewers.

We want thanks:

- Comunidad Autónoma de la Región de Murcia (with the support of FUNDACIÓN SÉNECA),
- Departamento de Matemática Aplicada y Estadística of Universidad Politécnica de Cartagena.
- Departamento de Matemáticas of Universidad de Murcia,
- Ayuntamiento de Lorca.
- Ayuntamiento de Cartagena.

CMMSE 2008 Plenary Speakers

- Erikki Brändas, Uppsala Universitet, Uppsala, Sweden
- Víctor Jiménez, Dept Matemáticas Universidad Murcia Spain
- Tatiana Levitina, *Institut Computational Mathematics, TU Braunschweig, Germany.*
- Shinnosuke Oharu, Chuo University, Tokyo, Japan
- J.I. Ramos Sobrados, Universidad Málaga. Spain
- Guido Van der Berghe, University Gent, Belgium
- Ezio Venturino Dipartimento di Matematica, University Torino, Italy

CMMSE 2008 Organising Committee

- Dr Juan L. G. Guirao (Chair, UPCT),
- Manuel Iván León Torres (UPCT) ,
- M. Araceli Queiruga Dios (USAL),
- M. Teresa de Bustos (USAL)

Contents:

Volume I

Contents:

Volume II

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Some results about inverse-positive matrices

Manuel F. Abad[1], **Maria T. Gassó**[1] **and Juan R. Torregrosa**[1]

[1] *Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia*

emails: maabrod@mat.upv.es, mgasso@mat.upv.es, jrtorre@mat.upv.es

Abstract

A nonsingular real matrix A is said to be inverse-positive if all the elements of its inverse are nonnegative. This class of matrices contains the M-matrices, from which inherit some of their properties and applications, especially in Economy. In this paper we present some new characterizations for inverse-positive matrices and we analyze when this concept is preserved by the sub-direct sum of matrices. We also study the inverse-positivity of real square matrices whose entries have a particular sign patterns.

Key words: Inverse-positive matrix, M-matrix, Sub-direct sum, Checkerboard pattern.
MSC 2000: 15A09

1 Introduction

In economics as well as other sciences, the inverse-positivity of real square matrices has been an important topic. A nonsingular real matrix $A = (a_{ij})$ is said to be *inverse-positive* if all the elements of its inverse are nonnegative. An inverse-positive matrix being also a Z-matrix, is a nonsingular M-matrix, so the class of inverse-positive matrices contains the nonsingular M-matrices, which have been widely studied and whose applications, for example, in iterative methods, dynamic systems, economics, mathematical programming, etc, are well known.

Of course, every inverse-positive matrix is not an M-matrix. For instance,

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

is an inverse-positive matrix that is not an M-matrix.

The inverse-positivity is preserved by multiplication, left or right positive diagonal multiplication, positive diagonal similarity and permutation similarity.

The paper is structured as follows. In section 2 we present some conditions in order to obtain new characterizations for inverse-positive matrices.

The sub-direct sum of matrices is a generalization of the usual sum of matrices. This concept was introduced by Johnson and Fallat in [2] and arises naturally in a variety of ways such as in matrix completion, overlapping subdomains in domain decomposition methods, etc. It also appears in many variants of additive Schwartz preconditioning, and when analyzing additive Schwartz methods for Markov chains. In section 3 we study the question of when the sub-direct sum of two inverse-positive matrices is an inverse-positive matrix.

In section 4 we analyze the inverse-positivity of matrices that very often occur in relation to Leontief model of circulating capital without joint production. For instance, matrices that for size 5×5 have the form

$$A = \begin{pmatrix} 1 & -a & 1 & -a & 1 \\ 1 & 1 & -a & 1 & -a \\ -a & 1 & 1 & -a & 1 \\ 1 & -a & 1 & 1 & -a \\ -a & 1 & -a & 1 & 1 \end{pmatrix},$$

where a is a real parameter with economic interpretation.

Johnson in [4] studied the possible sign patterns of a matrix which are compatible with inverse-positiveness. Following his results we analyze, in section 5, the mentioned concept for a particular type of pattern: the checkerboard pattern. We study the inverse-positivity of bidiagonal, tridiagonal and lower (upper) triangular matrices with checkerboard pattern. We obtain characterizations of the inverse-positivity for each class of matrices.

2 Characterization of inverse-positive matrices

The problem of characterizing inverse-positive matrices has been extensively dealt with in the literature (see for example [1]). In this section we show the main (and equivalent) characterizations of inverse-positive matrices.

Let x be a vector in R^n. In this paper we denote by $x \geq 0$ when all the components of x are nonnegative, $x > 0$ when all the components of x are nonnegative but not all zero simultaneously, and $x >> 0$ when all the components x are positive.

Let A be an $n \times n$ real matrix. Consider the nonempty subset S of $N = \{1, 2, \ldots, n\}$ and $T = S^c$, the complement of S with respect N. Now, we consider the following property

$$\text{If, for } x \geq 0, \text{we have } (Ax)_j > 0, \forall j \in S \text{ and } (Ax)_j = 0, \forall j \in T, \text{ then } x >> 0. \quad (1)$$

Let $\mathcal{C}$ be the class of $n \times n$ real matrices that satisfy property (1). We can establish the following result.

Proposition 2.1 *Let A be a nonsingular real matrix of size $n \times n$. Then $A \in \mathcal{C}$ if and only if $(A^{-1})_j > 0$, $\forall j \in S$ and given $i \in N$, $\exists j \in S$ such that $(A^{-1})_{ij} > 0$.*

Now, the main characterization of inverse-positive matrices is:

Theorem 2.1 *An $n \times n$ real matrix A is inverse-positive if and only if $\forall\ b >> 0$, $\exists\ x >> 0$ such that $Ax = b$.*

From this theorem we can establish the following result.

Corollary 2.1 *Let A be an $n \times n$ real matrix. Then A is inverse- positive if and only if A is monotone.*

We recall that an $n \times n$ real matrix is said to be *monotone* if $\forall x \in R^n$, if $Ax \geq 0$ then $x \geq 0$.

3 Sub-direct sum of inverse-positive matrices

As we have said in the introduction, the sub-direct sum was introduced by Johnson and Fallat in [2], where many of their properties were analyzed.

Let A, B be square matrices of n_1 and n_2 orders, respectively, and let k be an integer such that $1 \leq k \leq \min(n_1, n_2)$. Suppose that A and B are partitioned as follows

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where A_{22} and B_{11} are square matrices of order k. Then, the *sub-direct sum of order k* of A and B, indicated by $C = A \oplus_k B$, is the matrix

$$C = \begin{pmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} + B_{11} & B_{12} \\ 0 & B_{21} & B_{22} \end{pmatrix}.$$

Sub-direct sum of two inverse-positive matrices is not in general a inverse-positive matrix, as we can see in the following example.

Example 1 Let us consider the inverse-positive matrices

$$A = \left(\begin{array}{rr|rr} -1 & 2 & 0 & 0 \\ 3 & -1 & 0 & 0 \\ \hline -1 & -1 & 6 & -4 \\ -1 & -1 & -1 & 1 \end{array}\right), \qquad B = \left(\begin{array}{rr|rr} -2 & 1 & 0 & 0 \\ 8 & -1 & 0 & 0 \\ \hline -1 & -1 & -1 & 2 \\ -1 & -1 & 3 & -1 \end{array}\right).$$

It is easy to see that matrix

$$C = A \oplus_2 B = \left(\begin{array}{rr|rr|rr} -1 & 2 & 0 & 0 & 0 & 0 \\ 3 & -1 & 0 & 0 & 0 & 0 \\ \hline -1 & -1 & 4 & -3 & 0 & 0 \\ -1 & -1 & 7 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & -1 & -1 & 2 \\ 0 & 0 & -1 & -1 & 3 & -1 \end{array}\right)$$

is not inverse-positive.

In this paper we study the conditions for a sub-direct sum of inverse-positive matrices lies in the class, and it is appropriate to consider $k = 1$ and $k > 1$ separately.

We also study the question: if C is an inverse-positive matrix,

$$C = \begin{pmatrix} C_{11} & C_{12} & 0 \\ C_{21} & C_{22} & 0 \\ 0 & C_{32} & C_{33} \end{pmatrix}$$

may C be written as $C = A \oplus B$, such that A and B lie in the class when C_{22} is 1×1 and separately when C_{22} is $k \times k$ with $k > 1$. We obtain the following results.

1. Let A,B be the inverse-positive matrices

   $$A = \begin{pmatrix} A_{11} & 0 \\ a_{21}^T & a_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} b_{22} & 0 \\ b_{32}^T & B_{33} \end{pmatrix}.$$

   It can be proved that $C = A \oplus_1 B$ is inverse-positive.

   Moreover, let C be the inverse-positive matrix

   $$C = \begin{pmatrix} C_{11} & 0 & 0 \\ c_{21}^T & c_{22} & 0 \\ 0 & c_{32}^T & C_{33} \end{pmatrix}$$

   then C can always be expressed as $C = A \oplus_1 B$ where

   $$A = \begin{pmatrix} A_{11} & 0 \\ a_{21}^T & a_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} b_{22} & 0 \\ b_{32}^T & B_{33} \end{pmatrix},$$

   are inverse-positive matrices.

2. Let us now consider the inverse-positive matrices

   $$A = \begin{pmatrix} A_{11} & a_{12} \\ a_{21}^T & a_{22} \end{pmatrix} \qquad B = \begin{pmatrix} b_{22} & b_{23} \\ b_{32}^T & B_{33} \end{pmatrix}$$

   where a_{22} and b_{22} are scalars. It can be proved that $C = A \oplus_1 B$ is inverse-positive if and only if A_{11} and B_{33} are inverse-positive matrices.

   Let us now consider the inverse-positive matrix

   $$C = \begin{pmatrix} C_{11} & c_{12} & 0 \\ c_{21}^T & c_{22} & c_{23} \\ 0 & c_{32}^T & C_{33} \end{pmatrix},$$

   where C_{11} and C_{33} are inverse-positive matrices. It can be proved that C can always be expressed as $C = A \oplus_1 B$, where

$$A = \begin{pmatrix} C_{11} & c_{12} \\ c_{21}^T & a_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} b_{22} & c_{23} \\ c_{32}^T & C_{33} \end{pmatrix},$$

with $a_{22} = c_{22} - c_{23}C_{33}^{-1}c_{32}^T - \varepsilon$ and $b_{22} = c_{22} - a_{22}$

3. Using the standard patterns of A and B in order to study the properties of the sub-direct sum, it can be proved that, when A and B, both inverse-positive matrices, have the form

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix},$$

then $C = A \oplus_k B$ is inverse-positive.

Let us now consider the inverse-positive matrix

$$C = \begin{pmatrix} C_{11} & 0 & 0 \\ C_{21} & C_{22} & C_{23} \\ 0 & 0 & C_{33} \end{pmatrix}$$

It can be proved that C can always be expressed as $C = A \oplus_k B$, where

$$A = \begin{pmatrix} C_{11} & 0 \\ C_{21} & A_{22} \end{pmatrix} \qquad B = \begin{pmatrix} B_{22} & C_{23} \\ 0 & C_{33} \end{pmatrix}$$

are inverse-positive matrices.

4. Nevertheless, when A and B, both inverse-positive, are in the form:

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix},$$

let us designate $H = A_{22}^{-1} + B_{11}^{-1}$. It can be proved that if H is inverse-positive then the sub-direct sum of A and B is inverse-positive. The reciprocal does not hold.

Carrying on the same case, let us now designate $H' = A_{22} + B_{11}$. Given $b = [b_1, b_2, b_3]^T \in R_+^{n_1+n_2-k}$, with $b_1 \in R^{n_1-k}$, $b_2 \in R^k$ and $b_3 \in R^{n_2-k}$, we wonder if exists some vector $u > 0$ such as $Cu = b$. As A and B are inverse-positive, it exists the positive vectors $x_1 \in R^{n_1-k}$, $y_1 \in R^k$, $x_2 \in R^k$ and $y_2 \in R^{n_2-k}$, such as

$$A \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \qquad B \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} b_2 \\ b_3 \end{pmatrix}.$$

It can be proved that if matrix V

$$V = \begin{pmatrix} B_{21} & B_{21}x_2 \\ H' & A_{22}y_1 \end{pmatrix},$$

has rank k, then the sub-direct sum is inverse-positive.

Let us now consider the inverse-positive matrix

$$C = \begin{pmatrix} C_{11} & 0 & 0 \\ C_{21} & C_{22} & 0 \\ 0 & C_{32} & C_{33} \end{pmatrix}$$

It can be proved that C can always be expressed as $C = A \oplus_k B$, where

$$A = \begin{pmatrix} C_{11} & 0 \\ C_{21} & A_{22} \end{pmatrix} \qquad B = \begin{pmatrix} B_{22} & 0 \\ C_{32} & C_{33} \end{pmatrix}$$

are inverse-positive matrices.

Unlike, the general case $C = A \oplus_k B$, where A and B have the form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

and the conversely problem are still open problems.

4 Special pattern of inverse-positive matrices

In this section we are going to analyze the inverse-positivity of a particular type of square matrices. As we can see in [3], these matrices are used in Input-Output Leontief models with fixed capital. It is important to know in that models when these matrices are inverse-positive.

For example, for the particular 3×3 case, this special class of matrices have the form:

$$A = \begin{pmatrix} 1 & -a & 1 \\ 1 & 1 & -a \\ -a & 1 & 1 \end{pmatrix}.$$

It is easy to see that the inverse of A is

$$A^{-1} = \begin{pmatrix} \dfrac{-1}{a^2-a-2} & \dfrac{-1}{a^2-a-2} & \dfrac{-a+1}{a^2-a-2} \\ \dfrac{-a+1}{a^2-a-2} & \dfrac{-1}{a^2-a-2} & \dfrac{-1}{a^2-a-2} \\ \dfrac{-1}{a^2-a-2} & \dfrac{-a+1}{a^2-a-2} & \dfrac{-1}{a^2-a-2} \end{pmatrix}$$

when $a \neq -1$ or $a \neq 2$. This inverse is nonnegative if $1 \leq a \leq 2$.

It is easy to generalize the above example to real square matrices with an arbitrary size. Let A be a real matrix of size $n \times n$ whose diagonal elements are all equal to 1, and in each row, shifting from the diagonal element to the right, $(-a)$ and 1 appear alternately (jumping back to the first element at the rightmost one).

Example 2 In this example, A_1 and A_2 are matrices of this class with sizes odd and even, respectively.

$$A_1 = \begin{pmatrix} 1 & -a & 1 & -a \\ -a & 1 & -a & 1 \\ 1 & -a & 1 & -a \\ -a & 1 & -a & 1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & -a & 1 & -a & 1 \\ 1 & 1 & -a & 1 & -a \\ -a & 1 & 1 & -a & 1 \\ 1 & -a & 1 & 1 & -a \\ -a & 1 & -a & 1 & 1 \end{pmatrix}$$

It is important to notice that if the number of columns and rows of A is even, then A is singular for every value of parameter a, so it is not inverse-positive matrix.

When the size of the matrix is odd we can establish the following result.

Theorem 4.1 *Let A be an $n \times n$ real matrix with n odd, $n = 2k+1$.Then A is inverse-positive matrix if and only if $1 \leq a < (1 + \frac{1}{k})$.*

In addition, each element of the inverse of A is a quotient whose denominator is $(ka^2 - a - (k+1))$, while its numerator is either $((k-1)a - k)$ or $(-a+1)$.

5 Checkerboard inverse-positive matrices.

Following the results obtained by Johnson in [4], we analyze in this section a special pattern of matrices known as *checkerboard pattern*.

Definition 5.1 *A real matrix $A = (a_{ij})$, of size $n \times n$, is said to have a checkerboard pattern if $sign(a_{ij}) = (-1)^{i+j}$, $i, j = 1, 2, \ldots, n$.*

For example, the matrices introduced in the previous section, with even size, have checkerboard pattern.

Consider a nonsingular matrix A with checkerboard pattern. We are going to analyze the inverse-positivity when A is bidiagonal, tridiagonal or lower (upper) triangular matrix.

If A is bidiagonal then A is inverse-positive, since in this case A is an M-matrix.

However, in general a tridiagonal matrix is not an inverse-positive matrix, as we can observe in the following example.

Example 3 The tridiagonal matrix

$$A = \begin{pmatrix} 1 & -1 & 0 \\ -2 & 1 & -3 \\ 0 & -4 & 1 \end{pmatrix}$$

is nonsingular with checkerboard pattern, but it is not an inverse-positive matrix.

In the following result we present sufficient conditions for a tridiagonal matrix to be inverse-positive.

Proposition 5.1 *Let A be an $n \times n$ tridiagonal nonsingular matrix with checkerboard pattern. If $detA[\alpha] \geq 0$, for all $\alpha \subseteq \{1, 2, \ldots, n\}$ and $|\alpha| \geq 2$, then A is an inverse-positive matrix.*

Finally, when A is a nonsingular lower (upper) triangular matrix, with checkerboard pattern, the nonnegativity of its inverse is not guaranteed.

Example 4 Let us consider the lower triangular matrix

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

It is easy to check that A is not inverse-positive.

Let G_A be the associated graph to a matrix A. The following condition is related with the alternate paths that appear in G_A.

Definition 5.2 *Let A be an $n \times n$ nonsingular matrix with checkerboard pattern. Then A satisfies the PP-Condition if, for every alternate path $\{(i,k),(k,j),(i,j)\}$, we have*

$$p(i,j) \leq p(i,k)p(k,j), \quad i \neq k \neq j,$$

where $p(i,j)$ denotes the element of position (i,j) of the matrix A.

We need the next lemma in order to get the main result for the inverse-positivity of this class of matrices.

Lemma 5.1 *Let A be an $n \times n$ nonsingular lower triangular matrix, with checkerboard pattern, that satisfies the PP-condition. Then*

$$sign(detA[i, i+1, \ldots, n|i-1, i, \ldots, n-1]) = (-1)^{n+i-1}, \quad i = 2, 3, \ldots, n-1, n.$$

If A is an upper triangular matrix, the thesis of above lemma is

$$sign(detA[i-1, i, \ldots, n-1|i, i+1, \ldots, n]) = (-1)^{n+i-1}, \quad i = 2, 3, \ldots, n-1, n.$$

Theorem 5.1 *Let A be an $n \times n$ nonsingular lower(upper) triangular matrix with checkerboard pattern, that satisfies the PP-condition. Then A is an inverse-positive matrix.*

References

[1] A. Berman, R. Plemmons, *Nonnegative matrices in the Mathematical Sciences*, Siam, 1994.

[2] S. Fallat, C. Johnson, *Sub-direct sums and positivity classes of matrices.* Linear Algebra and its Applications, 288 (1999), 149-173.

[3] T. Fujimoto, J.A. Silva and A. Villar, *A Generalization of Theorems on Inverse-Positive Matrices*, Kagawa University, 2003.

[4] Charles R. Johnson, *Sign Patterns of Inverse Nonnegative Matrices*, Linear Algebra and its Applications, 55:69-80 (1983)

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Mathematical Models for Biohazards and Farmer's Behavior in a Swine Epidemic in the Cuneo Area, Italy.

Davide Abbona[1], Bruno Sona[2] and Ezio Venturino[1]

[1] *Dipartimento di Matematica, via Carlo Alberto 10,,*
Universita' di Torino, 10123 Torino, Italia

[2] *Servizi Veterinari ASL CN1, Savigliano, Regione Piemonte, Italia.*

emails: , `vete.villafalletto@asl17.it`, `ezio.venturino@unito.it`

Abstract

We formulate and analyze a mathematical model for the spread of the Aujeszky disease, to determine possible strategies for its control. We also indirectly simulate inappropriate human interventions, which badly affect the disease spread.

Field data on the blood samples collected by the veterinarians for the serologic exam of breeding animals of the Cuneo province have been used to calibrate important model parameters. Birth and mortality rates partitioned among the disease-related and the disease-independent animals have thus been determined.

Starting from a well-known epidemic model for realistic situations we modified it to account for possible fluxes between susceptible, "vaccinated" and infected animals, incorporating also the biohazards. We also simulate the farmer's behavior who does not fully comply with correct vaccination policies. The outcomes of the analysis are interpreted in the light of possible epidemics containment policies.

Key words: epidemics, Aujeszky disease, biosafety, vaccination
MSC 2000: 92D30, 92D25

1 Introduction

The Aujeszky disease (A.D.) is caused by the Herpevirus 1 suis (ADV or SHV-1), affecting several wild and domestic species, but in particular hogs. It is not lethal, but it causes several disorders in the affected animals and ultimately it constitutes an economical burden for the farmer. Once contracted the disease, the infected animal cannot recover from it. Almost every country in Europe is affected. An eradication policy based on vaccination has been attempted in Italy in the past ten years, with mixed results.

Here we formulate and analyze a mathematical model possibly aimed at defining the realisation of a disease eradication plan at nonprohibitive costs. To determine strategies for the desirable if at all possible disease control, we begin by modeling the description of the evolution of the epidemics. A partial result has already been obtained, [6]. Noteworthy in the previous study is the modeling of the absence of biosafety measures, which may allow disease infiltration into a non affected farm by external vectors and not by direct contact between infected and susceptible animals. In this investigation we extend the previous model by taking into account the human intervention in the formulation, which if inappropriate favors the disease spread.

After presenting the picture of the situation and the sampling methods used in the next Section, we describe the mathematical model in Section 3 and perform its analysis for stable equilibria in Section 4. Section 5 contains a mathematical discussion of the results and some operative conclusions are drawn in the final Section.

2 Methods

This study is an attempt to investigate the situation using mathematical methods, in strict collaboration with the veterinarians studying the disease on the field, [1]. We have considered breeding farms in the area of the towns Villafalletto and Vottignasco in the Cuneo province in Piedmont, NW Italy. This part of the region is considered as a single giant epidemiological unit, since the swine density exceeds 3200 units per Km^2 with a total number of 90000 units.

Blood sample data collected according to the law (D.M. April 1st, 1997) in the period 1997-2004 for the serologic exam for A.D. constitute the basis for our analysis. From these and for each breeding farm birth rates have been determined together with natural and disease-related mortalities. These informations have been used in the model to give reasonably accurate numeric values to the relevant parameters, in order to validate the subsequent analysis and simulations.

From the mathematical point of view, in [6] we started from a well-known and accepted epidemics model for realistic situations, [2]. But contrary to the assumptions of the classical epidemiological model, [5], we allowed the total population to reproduce, as done in more recent models for disease spread, [3, 4]. The model studied in [6] has been here modified to include one more important feature. Based on the fact that disease prevalence went down after the first three years of law implementation, to come up again in the years 2000-2004, it is indeed argued whether the vaccination is in the end at all useful. This may be due to an intrinsic weakness of the vaccine, or to the bad implementation on the part of the farmers. The vaccine should indeed be administered three times in the lifetime of the animal, the first two times in the first months of life, the third one when the animal is about one year and a half old. Since by that time the farmer is in general ready to sell it, it may happen that the farmer avoids to administer the third vaccination, maybe thinking it is then unnecessary and to save on its costs. Whatever the cause, in any case the distinguishing feature of this model is represented by the introduction of the class of animals on which the vaccine is ineffective, allowing possible fluxes between susceptible, "vaccinated" and infected animals. We discuss the outcomes of our analysis in terms of possible policies to contain the epidemics.

3 The model

In formulating the model we take into account the following basic variables. First of all we consider the susceptible animals, $S(t)$. The latter is then subdivided into two further classes, the class of "formally" vaccinated animals, $V(t)$ and the one of susceptible or unvaccinated animals, $U(t)$. Then there is the infected class, $I(t)$. We allow transitions among these classes assuming that the vaccine does not always have a full effect, or is not implemented correctly. The transitions are described in the following equations. Notice that the disease is unrecoverable, so that once infected, an animal carries it for life, no transition back from the class I to either U or V is allowed.

We also assume that all individuals reproduce and newborns at birth are susceptible, due to some immunization gotten from the mother, there is no possibility of vertical transmission of the disease. Of course they will lose this immunity as they grow older. They then all belong to class U or V at birth, with respective birth rates ρ_U and ρ_V.

We consider then the following model

$$\begin{aligned} \dot{U} &= \rho_U N - \mu_S U - \beta_U \frac{UI}{N} - \tau U + \alpha V - \sigma U, \qquad (1) \\ \dot{I} &= \beta_U \frac{UI}{N} + \beta_V \frac{VI}{N} + \tau U + \tau V - \mu_I I, \\ \dot{V} &= \rho_V N - \mu_S V - \beta_V \frac{VI}{N} - \tau V - \alpha V + \sigma U, \end{aligned}$$

in addition to

$$N = U + V + I. \qquad (2)$$

The first equation says that all newborns coming from parents of whichever class are born sound, i.e. they are susceptible to the disease. The class of susceptibles is subject to natural mortality μ_S. Some of its members migrate to the class V of the vaccinated at rate σ, but either for ineffectiveness of the vaccine or faults in its implementation we assume they can migrate back from class V into the susceptibles at rate α. This is clearly modeled by the last two terms of the first equation (1). The disease affects them via the incidence β, which expresses contagion of a susceptible upon direct contact with an infected animal, while the parameter τ expresses the fact that the susceptible can get infected also by other means, by vectors carried into its environment by external factors. Thus in a sense this parameter models the biohazards, which should be tackled by suitable biosafety measures in the farm.

The second equation gives the dynamics of infected individuals. They enter class I via direct contact at the rate β specified above, or via an external vector at rate τ; in both cases they can enter class I either from class U or from class V if the vaccine is ineffective. Finally they are subject to disease-related mortality μ_I.

The third equation (1) describes the evolution of the vaccinated class. The first term again represents the newborns, then there is the natural mortality term, the disease incidence which may affect also these animals, if the vaccine is not so effective, again the infection caused by external factors, and finally the vaccination at rate σ and the loss of immunization at rate α.

The available parameter values from veterinarians field measuments are $N = 90000$ for the total hogs population in the epidemiological unit, $\mu_S = 0.084$ represents the average natural mortality, $\mu_I = 0.087$ is the average disease-related mortality, $\rho = 0.107$ represents the mean birth rate in the whole epidemiological unit. Since the field measurements provide only a lump natality rate ρ, we assume the birth rates to be partitioned as follows

$$\rho_U = \rho\frac{U}{N}, \quad \rho_V = \rho\frac{V}{N}. \tag{3}$$

The remaining parameter description is as mentioned above, with β_U and β_V denoting the horizontal disease incidences respectively for the classes of susceptibles and vaccinated, τ the biohazards, α the loss of immunity due to failure in the vaccine or in its administration: it is thus a migration rate into the class of susceptibles from the class of vaccinated. Finally σ is the vaccination rate, expressing a migration from U into V.

We now introduce three new variables, given by the subpopulations fractions u,i e v, which in view of the constraint on the total population (2) are clearly related to each other as follows

$$u = \frac{U}{N}, \quad i = \frac{I}{N}, \quad v = \frac{V}{N}, \quad u + v + i = 1. \tag{4}$$

Notice that the position (4) entails that upon differentiation for instance

$$\dot{u} = \frac{\dot{U}}{N} - \frac{U}{N}\frac{\dot{N}}{N} = \frac{\dot{U}}{N} - \frac{U}{N}\left[\frac{\dot{U}}{N} + \frac{\dot{V}}{N} + \frac{\dot{I}}{N}\right], \tag{5}$$

and similarly for the other fractions. Thus substituting from (1) we obtain

$$\begin{aligned} \dot{u} &= \mu_S u^2 + ui(\mu_I - \beta_U) + \rho_U + \mu_S uv + \alpha v - u(\mu_S + \tau + \sigma + \rho_U + \rho_V), \\ \dot{v} &= \mu_S v^2 - v(\mu_S + \tau + \alpha + \rho_U + \rho_V) + vi(\mu_I - \beta_V) + \rho_V + \sigma u + \mu_S vu, \\ \dot{i} &= \mu_I i^2 + ui(\mu_S + \beta_U) + vi(\beta_V + \mu_S) - i(\mu_I + \rho_V + \rho_U) + \tau u + \tau v. \end{aligned} \tag{6}$$

Then by using into (6) the assumption on the birth rates (3) we have

$$\begin{aligned} \dot{u} &= \mu_S u^2 - u(\mu_S + \tau + \sigma) + \mu_S vu + ui(\mu_I - \beta_U + \rho) + \alpha v, \\ \dot{v} &= \mu_S v^2 + vi(\rho + \mu_I - \beta_V) + \mu_S uv + \sigma u - v(\mu_S + \tau + \alpha), \\ \dot{i} &= i^2(\mu_I + \rho) - i(\mu_I + \rho + \tau) + ui(\beta_U + \mu_S) + vi(\beta_V + \mu_S) + \tau. \end{aligned} \tag{7}$$

Finally on eliminating the variable i from (4) we have the reduced model description via the equations

$$\begin{aligned} \dot{u} &= (\mu_S - \mu_I - \rho + \beta_U)u^2 + uv(\mu_S - \mu_I - \rho + \beta_U) - u(\mu_S + \tau + \sigma - \rho - \mu_I + \beta_U) \\ &\quad + \alpha v, \\ \dot{v} &= v^2(\mu_S - \rho - \mu_I + \beta_V) + uv(\mu_S - \rho - \mu_I + \beta_V) + v(\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V) \\ &\quad + \sigma u \end{aligned} \tag{8}$$

In view of the constraint (4), we seek the solutions of (8) in the unit simplex $\Omega \equiv \{(u, v) \in \mathbf{R}^2 : 0 \leq u, v \leq 1\}$.

4 Analysis

Let us now seek the system equilibria. By equating to zero the right hand sides of (8) we find the following equations

$$(\mu_S-\mu_I-\rho+\beta_U)u(u+v)+\alpha v-u(\mu_S+\tau+\sigma-\rho-\mu_I+\beta_U)=0, \tag{9}$$

$$(\mu_S-\rho-\mu_I+\beta_V)v(v+u)+\sigma u+v(\rho-\mu_S-\tau-\alpha+\mu_I-\beta_V)=0. \tag{10}$$

These represent conic sections. To study the first conic, we consider its invariants

$$\Delta_1 \equiv \begin{vmatrix} (\Delta_1)_{1,1} & (\Delta_1)_{1,2} & (\Delta_1)_{1,3} \\ (\Delta_1)_{2,1} & 0 & (\Delta_1)_{2,3} \\ (\Delta_1)_{3,1} & (\Delta_1)_{3,2} & 0 \end{vmatrix}$$

with elements given by

$$(\Delta_1)_{1,1}=\mu_S-\rho+\beta_U-\mu_I, \quad (\Delta_1)_{1,2}=(\Delta_1)_{2,1}=\frac{\mu_S-\rho+\beta_U-\mu_I}{2},$$
$$(\Delta_1)_{1,3}=(\Delta_1)_{3,1}=\frac{\rho-\mu_S-\tau-\beta_U+\mu_I-\sigma}{2}, \quad (\Delta_1)_{2,3}=(\Delta_1)_{3,2}=\frac{\alpha}{2},$$

and

$$\delta_1 \equiv \begin{vmatrix} (\Delta_1)_{1,1} & (\Delta_1)_{1,2} \\ (\Delta_1)_{2,1} & 0 \end{vmatrix}.$$

Evaluating the determinants, then we find

$$\begin{aligned}
\Delta_1 &= \frac{\alpha}{4}[-\alpha(\mu_S-\rho+\beta_U-\mu_I)+(\mu_S-\rho-\mu_I+\beta_U)(\rho-\mu_S-\tau-\sigma+\mu_I+\beta_U)] \\
&= \frac{\alpha}{4}(\mu_S-\rho-\mu_I+\beta_U)\times(\rho-\mu_S-\tau-\sigma+\mu_I+\beta_U-\alpha), \\
\delta_1 &= -\frac{\alpha}{4}(\mu_S-\rho+\beta_U-\mu_I)^2.
\end{aligned}$$

Since $\delta_1<0$ it is therefore a hyperbola. Its center is the point (u^*,v^*) where

$$\begin{aligned}
u^* &\equiv \frac{1}{\delta_1}\begin{vmatrix} (\Delta_1)_{1,2} & (\Delta_1)_{1,3} \\ (\Delta_1)_{2,2} & (\Delta_1)_{2,3} \end{vmatrix} = -\frac{\alpha}{\mu_I-\mu_S-\beta_U+\rho}, \\
v^* &\equiv \frac{1}{\delta_1}\begin{vmatrix} (\Delta_1)_{1,3} & (\Delta_1)_{1,1} \\ (\Delta_1)_{2,3} & (\Delta_1)_{2,1} \end{vmatrix} = \frac{\mu_S+\tau+\sigma-\mu_I-\rho+\beta_U+2\alpha}{\mu_S-\rho-\mu_I+\beta_U}.
\end{aligned}$$

The asymptotes of this hyperbola are given by

$$\begin{aligned}
&(\mu_S-\rho+\beta_U-\mu_I)u^2+(\mu_S-\rho-\mu_I+\beta_U)uv \\
&+(\rho-\mu_S-\tau-\sigma+\mu_I-\beta_U)u+\alpha v-\frac{\Delta_1}{\delta_1}=0,
\end{aligned} \tag{11}$$

where

$$\frac{\Delta_1}{\delta_1}=\frac{\alpha^2-\alpha(\rho-\mu_S-\sigma+\mu_I-\beta_U)}{\mu_S-\rho-\mu_I+\beta_U}.$$

To find them explicitly, upon division of (11) by $\mu_S - \rho - \mu_I + \beta_U$, we find

$$T(u,v) = u^2 + \frac{\rho - \mu_S - \tau - \sigma + \mu_I - \beta_U}{\mu_S - \rho - \mu_I + \beta_U} u + vu + \frac{\alpha}{\mu_S - \rho - \mu_I + \beta_U} v \\ - \frac{\alpha^2 - \alpha(\rho - \tau - \mu_S - \sigma + \mu_I - \beta_U)}{(\mu_S - \rho - \mu_I + \beta_U)^2} = 0.$$

Let us assume $T(u,v)$ to be the product of two linear functions with undetermined coefficients, so that

$$T(u,v) = (\widetilde{A}u + \widetilde{B}v + \widetilde{C})(\widetilde{D}u + \widetilde{E}v + \widetilde{F}) = 0.$$

Upon equating coefficients of like powers, we find that the following equations must be satisfied by $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$, $\widetilde{E}$, $\widetilde{F}$,

$$\begin{array}{rcl} u^2 & : & \widetilde{A}\widetilde{D} = 1, \qquad uv : \ \widetilde{A}\widetilde{E} + \widetilde{B}\widetilde{D} = 1, \qquad v : \ \widetilde{B}\widetilde{F} + \widetilde{C}\widetilde{E} = \dfrac{\alpha}{(\mu_S - \rho - \mu_I + \beta_U)}, \\ u & : & \widetilde{A}\widetilde{F} + \widetilde{D}\widetilde{C} = \dfrac{(\rho - \tau - \mu_S - \sigma + \mu_I - \beta_U)}{(\mu_S - \rho - \mu_I + \beta_U)}, \qquad v^2 : \ \widetilde{B}\widetilde{E} = 0, \\ u^0 v^0 & : & \widetilde{F}\widetilde{C} = \dfrac{\alpha^2 - \alpha(\rho - \tau - \mu_S - \sigma + \mu_I - \beta_U)}{(\mu_S - \rho - \mu_I + \beta_U)^2}. \end{array} \tag{12}$$

Without loss of generality, taking for instance $\widetilde{E} = 0$, to satisfy the second above equation, we have then the straight lines in the form

$$u = -\frac{\widetilde{F}}{\widetilde{D}}, \quad v = -\frac{\widetilde{A}}{\widetilde{B}} u - \frac{\widetilde{C}}{\widetilde{B}}.$$

We find now their coefficients as follows. From the fifth and the second of (12), since $\widetilde{E} = 0$ we have

$$\frac{\widetilde{F}}{\widetilde{D}} \equiv \frac{\widetilde{F}\widetilde{B}}{\widetilde{D}\widetilde{B}} = -\frac{\alpha}{\mu_S + \beta_U - \mu_I - \rho}.$$

so that the first asymptote is

$$u = -\frac{\alpha}{\mu_S + \beta_U - \mu_I - \rho}. \tag{13}$$

Then the first two equations of (12) give

$$\frac{\widetilde{A}}{\widetilde{B}} \equiv \frac{\widetilde{A}\widetilde{D}}{\widetilde{B}\widetilde{D}} = 1$$

and the last two in turn yield

$$\frac{\widetilde{C}}{\widetilde{B}} \equiv \frac{\widetilde{F}\widetilde{C}}{\widetilde{B}\widetilde{F}} = \frac{\alpha - (\rho - \mu_S - \sigma - \tau + \mu_I - \beta_U)}{(\mu_S - \rho - \mu_I + \beta_U)}.$$

Thus the second asymptote is

$$v = -u + \frac{\alpha - (\rho - \mu_S - \sigma - \tau + \mu_I - \beta_U)}{(\mu_S - \rho - \mu_I + \beta_U)}. \tag{14}$$

The intersections with the coordinate axes of the hyperbola (9) are the origin and the points $v = 0$ and the roots of the quadratic

$$u^2(\mu_S - \rho + \beta_U - \mu_I) + u(\rho - \mu_S - \tau - \sigma + \mu_I - \beta_U) = 0$$

which are explicitly

$$u = 0, \quad u = -\frac{(\rho - \mu_S - \tau - \sigma + \mu_I - \beta_U)}{(\mu_S - \rho + \beta_U - \mu_I)}.$$

We study now the conic (10). Its invariants can be determined as follows.

$$\Delta_2 \equiv \begin{vmatrix} 0 & (\Delta_2)_{1,2} & (\Delta_2)_{1,3} \\ (\Delta_2)_{2,1} & (\Delta_2)_{2,2} & (\Delta_2)_{2,3} \\ (\Delta_2)_{3,1} & (\Delta_2)_{3,2} & 0 \end{vmatrix}$$

which has the elements

$$\begin{aligned} (\Delta_2)_{1,2} &= (\Delta_2)_{2,1} = \frac{\mu_S - \rho - \mu_I + \beta_V}{2}, \quad (\Delta_2)_{1,3} = (\Delta_2)_{3,1} = \frac{\sigma}{2}, \\ (\Delta_2)_{2,2} &= \mu_S - \rho - \mu_I + \beta_V, \quad (\Delta_2)_{2,3} = (\Delta_2)_{3,2} = \frac{\rho + \mu_I - \mu_S - \tau - \alpha - \beta_V}{2}. \end{aligned}$$

Upon evaluation, from these we find

$$\begin{aligned} \Delta_2 &= \frac{\sigma}{4}(\mu_S - \rho - \mu_I + \beta_V) \times (\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V - \sigma), \\ \delta_2 &\equiv \begin{vmatrix} 0 & (\Delta_2)_{1,2} \\ (\Delta_2)_{2,1} & (\Delta_2)_{2,2} \end{vmatrix} = -\frac{(\mu_S - \rho + \beta_V - \mu_I)^2}{4}. \end{aligned}$$

Again $\delta_2 < 0$ shows that also (10) is a hyperbola. To find its center we use once again the invariant method

$$\begin{aligned} u_2 &\equiv \frac{1}{\delta_2}\begin{vmatrix} (\Delta_2)_{1,2} & (\Delta_2)_{1,3} \\ (\Delta_2)_{2,2} & (\Delta_2)_{2,3} \end{vmatrix} = -\frac{-(\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V) + 2\sigma}{\mu_S - \rho - \mu_I + \beta_V}, \\ v_2 &\equiv \frac{1}{\delta_2}\begin{vmatrix} (\Delta_2)_{1,3} & 0 \\ (\Delta_2)_{2,3} & (\Delta_2)_{2,1} \end{vmatrix} = -\frac{\sigma}{\mu_S - \rho - \mu_I + \beta_V}. \end{aligned}$$

The asymptotes are found from

$$\begin{aligned} &(\mu_S - \rho - \mu_I + \beta_V)v^2 + (\mu_S - \rho - \mu_I + \beta_V)uv \\ &\quad + (\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V)v + \sigma u = \frac{\Delta_2}{\delta_2}, \end{aligned}$$

where

$$\frac{\Delta_2}{\delta_2} = -\frac{\sigma(\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V) + \sigma^2}{(\mu_S - \rho - \mu_I + \beta_V)}.$$

To explicitly determine the asymptotes, upon division by $\mu_S - \rho - \mu_I - \beta_V$ we have

$$Q(u,v) = v^2 + \frac{(\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V)}{(\mu_S - \rho - \mu_I + \beta_V)}v + uv + \frac{\sigma}{(\mu_S - \rho - \mu_I + \beta_V)}u + \frac{\sigma(\rho - \mu_S - \tau - \alpha + \mu_I - \beta_V) + \sigma^2}{(\mu_S - \rho - \mu_I + \beta_V)^2} = 0.$$

Again let us take $Q(u,v)$ in factored form

$$Q(u,v) \equiv (Gu + Hv + I)(Lu + Mv + N) = 0.$$

Equating like powers of the variables, we thus find

$$\begin{array}{llllll} u^2: & GL = 0, & uv: & GM + HL = 1, & u: & GN + IL = \dfrac{\sigma}{A}, \\ v: & HN + IM = \dfrac{B}{A}, & v^2: & HM = 1, & u^0v^0: & IN = \dfrac{\sigma E}{A^2}. \end{array} \tag{15}$$

In these equations we have used the following shorthands

$$\begin{array}{rcl} C & = & \mu_S - \rho - \mu_I = -0.110 < 0, \quad A = C + \beta_V, \quad D = C + \beta_U, \quad \Gamma = C + \tau, \\ B & = & \Gamma + \alpha + \beta_V = A + \alpha, \quad E = B - \sigma, \quad F = \Gamma + \sigma + \beta_U = D + \sigma. \end{array} \tag{16}$$

In this way $Q(u,v)$ can be rewritten as

$$Q(u,v) = v^2 + \frac{B}{A}v + uv + \frac{\sigma}{A}u + \frac{\sigma E}{A^2}.$$

Let us take $G = 0$, without loss of generality. We have then the asymptotes

$$v = -\frac{I}{H}, \quad v = -\frac{L}{M}u - \frac{N}{M}.$$

The second and third equations (15) give

$$\frac{I}{H} = \frac{IL}{HL} = \frac{\sigma}{A}.$$

From the second and fifth equation (15) we have

$$\frac{L}{M} = \frac{HL}{HM} = 1.$$

Also using the second, the sixth, the fifth and the third equation (15) we have

$$\frac{N}{M} = \frac{HL}{HM}\frac{IN}{IL} = \frac{E}{A}.$$

The asymptotes have then the equations

$$v = -u + \frac{E}{A}, \quad v = -\frac{\sigma}{A}. \tag{17}$$

5 Discussion

Let us now introduce new notations, to simplify the subsequent discussion. Using (16) the conics can be rewritten as

$$Du^2 - Fu + Duv + \alpha v = 0 \tag{18}$$
$$Av^2 - Bv + Auv + \sigma u = 0 \tag{19}$$

The center of (18) becomes

$$u_0 = -\frac{\alpha}{D}, \quad v_0 = 1 + \frac{\tau + \sigma + 2\alpha}{D}.$$

Set

$$u_* = 1 + \frac{\tau + \sigma}{D}, \quad v_0 = u_* - 2u_0,$$

its axes intersections are then the origin and $(u_*, 0)$. Its asymptotes are finally

$$u = u_0, \quad v = -u + 1 + \frac{\tau + \sigma - \alpha}{C + \beta_U} = -u + u_* + u_0.$$

Studying the intersection with the straight line $u + v = 1$ we find

$$\bar{u} = \frac{\alpha}{\tau + \alpha + \sigma} < 1, \quad \bar{v} = \frac{\tau + \sigma}{\tau + \alpha + \sigma} < 1.$$

Set now

$$\tilde{v}_* = \frac{\tau + \alpha + 2\sigma}{A}.$$

The center of (19) becomes

$$\tilde{u}_0 = 1 + \frac{\tau + \alpha + 2\sigma}{A} = \tilde{v}_* - 2\tilde{v}_0, \quad \tilde{v}_0 = -\frac{\sigma}{A},$$

its axes intersections are then the origin and $(0, \tilde{v}_*)$. Its asymptotes are finally

$$v = \tilde{v}_0, \quad v = -u - \tilde{v}_* + \tilde{v}_0.$$

Studying the intersection with $u + v = 1$ we have

$$\tilde{v} = \frac{\sigma}{\tau + \alpha + \sigma}, \quad \tilde{u} = \frac{\tau + \alpha}{\tau + \alpha + \sigma}.$$

To study the flow in the unit simplex Ω we need to determine the mutual positions of the hyperbolae (18) and (19). To do this, we can use the above informations and discriminate between their slopes at the origin. In particular we find that the slope at the origin of (18) is larger than the one of (19) if the following inequalities are satisfied

$$C + \beta_V + \tau + \alpha > 0, \quad C + \beta_U + \tau + \sigma > \frac{\alpha\sigma}{C + \beta_V + \tau + \alpha} > 0, \tag{20}$$

or

$$B < 0, \quad F > \frac{\alpha\sigma}{B}, \quad \frac{\alpha\sigma}{B} < 0. \tag{21}$$

Conversely the slope at the origin of (19) is larger than the one of (18) if the following inequalities are satisfied

$$B < 0, \quad F < \frac{\alpha\sigma}{B} < 0, \tag{22}$$

or

$$B > 0 \quad F < \frac{\alpha\sigma}{B}, \quad \frac{\alpha\sigma}{B} > 0. \tag{23}$$

On the border $u + v = 1$ of the unit simplex Ω in the uv phase plane the flow is directed upwards (increasing v) if

$$u < u^{\dagger} \equiv \frac{1}{2} \frac{\tau - 2\alpha}{\tau + \sigma - \alpha}. \tag{24}$$

Notice that $u^{\dagger} < \frac{1}{2}$ if and only if $\tau + \sigma > \alpha$. The point $(u^{\dagger}, 1 - u^{\dagger})$ on the line $u + v = 1$ represents thus a saddle. Above it the flow goes upwards, below it goes downwards.

We need finally to determine the flow inside the unit simplex in the uv phase plane. For this the informations on the two conics (18) and (19) need to be merged. Several pictures can be drawn corresponding to several cases of possible intersections among the two curves and positions of the other relevant points on the coordinate axes.

6 Conclusions

We now summarize the ultimate behavior of the system, identifying when possible its ω-limit points. In some cases the system trajectories naturally evolve toward the line $u + v = 1$, corresponding to the disease-free environment, i.e. $i = 0$ in the three dimensional phase space uvi, recalling (4). These are the equilibria we should strive for. Some instances in which they are found are as discussed here below and some other situations leading to the same final outcome are listed in the Table.

For $D > 0$, $A > 0$, $\tilde{u}_0 > 0$, $\tilde{v}_0 > 1$ and $u_* > 1$ cases (20) or (21) give an internal saddle point, the origin is a stable equilibrium, implying that $i = 1$ i.e. a pandemia affecting the whole population. There is also a stable equilibrium on the line $i = 0$. It is enough that trajectories lie in its basin of attraction for the disease to vanish.

An endemic stable equilibrium is found instead for the case $C + \beta_V < 0$, $C + \beta_U > 0$ with $\tilde{u} > \bar{u}$. But in the same situation instead with $\tilde{u} < \bar{u}$ the stable equilibrium moves on the line $u + v = 1$.

In all possible cases, the above result show that it would be possible to eradicate the epidemics by acting appropriately on the relevant parameters of the model, so as to satisfy the conditions leading to stable disease-free equilibria. Moreover there is also the possibility of choosing which parameters to act upon, so that the above inequalities are satisfied. This allows some freedom for the policy maker in the choice

Table: Parameter combinations possibly leading to disease eradication.

$D > 0$	$A < 0$	$\tilde{v}_0 > 0$	$0 < \tilde{v}_* < \tilde{v}_0$	$\tilde{u} < \bar{u}$	
$D > 0$	$A < 0$	$\tilde{v}_0 < 0$	$\tilde{v}_* < 0 < \tilde{v}_0$	$\tilde{u} < \bar{u}$	
$D > 0$	$A < 0$	$\tilde{v}_0 > 0$	$1 < \tilde{v}_*$	$0 > \tilde{v}_0$	$\tilde{u} < \bar{u}$
$D < 0$	$A < 0$	$\tilde{u} < \bar{u}$			
$D < 0$	$A < 0$	$\tilde{u} > \bar{u}$			
$D > 0$	$A > 0$	$1 < \tilde{v}_*$	$\tilde{u} < \bar{u}$		
$D < 0$	$A < 0$	$\tilde{u}_0 > 0$	$1 < \tilde{u}_*$	$\tilde{u} < \bar{u}$	
$D < 0$	$A > 0$	$\tilde{v}_0 > 0$	$\tilde{u} < \bar{u}$		
$D < 0$	$A < 0$	$\tilde{u}_0 < 0$	$\tilde{v}_* < 0 < \tilde{v}_0$	$\tilde{u} > \bar{u}$	

of the most appropriate means of fighting the epidemics. In particular there would be the possibility of better enforcing the vaccination program, so as to augment σ and at the same time decrease α, or rather to act on preventive measures, such as to counteract the biohazards which are prone to spread the disease horizontally. This can be implemented by taking suitable biosafety restrictions, so as to diminish the disease incidence β_U and β_V, and also to reduce the possibility of importing the disease through external vectors, thus obtaining a smaller τ.

References

[1] M. Drigo, M. Dalla Pozza, G. Ferrari, M. Martini, B. Sona, A study of progress of Aujeszky 's disease control programme in Italy using survival analysis, *J. Vet. Med. B*, 50: 191–195, 2003.

[2] H.C. Davison, N.P. French, D. Clancy, A.J. Trees, Mathematical models of *neospora caninum* infection in dairy cattle: transmission and options for control. *Int. J. for Parasitology*, 29: 1691–1704, 1999.

[3] L. Q. Gao and H. W. Hethcote, Disease trasmission models with density-dependent demographics, *J. Math. Biol.*, 30: 717–731, 1992.

[4] J. Mena-Lorca and H. W. Hethcote, Dynamic models of infectious diseases as regulators of population sizes, *J. Math. Biol.*, 30: 693–713, 1992.

[5] W. O. Kermack and A. G. Mc Kendrick, Contributions to the mathematical theory of epidemics, part 1,. *Proc. Roy. Soc. London Ser. A*, 115: 700–721, 1927.

[6] D. Abbona, M. Drigo, M. Martini, B. Sona, E. Venturino, Modellizzazione matematica della malattia di Aujeszky in un'area ad altissima densita' di allevamenti suini della provincia di Cuneo, (Modelling mathematically the Aujeszky disease in a swine densely populated area of the Cuneo province, Italy). *Atti XXXII Convegno SIPAS*, 219–231, 2006.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Torus bifurcations, isolas and chaotic attractors in a simple dengue model with ADE and temporary cross immunity

Maíra Aguiar[1], **Nico Stollenwerk**[1] **and Bob W. Kooi**[2]

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

[2] *Faculty of Earth and Life Sciences, Department of Theoretical Biology, Vrije Universiteit Amsterdam, Nederland*

emails: maira@igc.gulbenkian.pt, nico@ptmat.fc.ul.pt, kooi@falw.vu.nl

Abstract

We analyse an epidemiological model of competing strains of pathogens and hence differences in transmission for first versus secondary infection due to interaction of the strains with previously aquired immunities, as has been described for dengue fever (in dengue known as antibody dependent enhancement, ADE). Such models show a rich variety of dynamics through bifurcations up to deterministic chaos. Including temporary cross-immunity even enlarges the parameter range of such chaotic attractors, and also gives rise to various coexisting attractors, which are difficult to identify by standard numerical bifurcation programs using continuation methods. A combination of techniques, including classical bifurcation plots and Lyapunov exponent spectra has to be applied in comparison to get further insight into such dynamical structures. Here we present for the first time multi-parameter studies in a range of biologically plausible values for dengue. The multi-strain interaction with the immune system is expected to also have implications for the epidemiology of other diseases.

Key words: numerical bifurcation analysis, Lyapunov exponents, $\mathbb{Z}_2$ symmetry, coexisting attractors, antibody dependent enhancement (ADE)

1 Introduction

Epidemic models are classically phrased in ordinary differential equation (ODE) systems for the host population divided in classes of susceptible individuals and infected ones (SIS system), or in addition, a class of recovered individuals due to immunity after an infection to the respective pathogen (SIR epidemics). The infection term includes a product of two variables, hence a non-linearity which in extended systems can cause complicated dynamics. Though these simple SIS and SIR models only show fixed points as equilibrium solutions, they already show non-trivial equilibria arising from bifurcations, and in stochastic versions of the system critical fluctuations at the threshold. Further refinements of the SIR model in terms of external forcing or distinction of infections with different strains of a pathogen, hence classes of infected with one or another strain recovered from one or another strain, infected with more than one strain etc., can induce more complicated dynamical attractors including equilibria, limit cycles, tori and chaotic attractors.

Classical examples of chaos in epidemiological models are childhood diseases with extremely high infection rates, so that a moderate seasonal forcing can generate Feigenbaum sequences of period doubling bifurcations into chaos. The success in analysing childhood diseases in terms of modelling and data comparison lies in the fact that they are just childhood diseases with such high infectivity. Otherwise host populations cannot sustain the respective pathogens. In other infectious diseases much lower forces of infection have to be considered leading to further conceptual problems with noise affecting the system more than the deterministic part, leading even to critical fluctuations with power law behaviour, when considering evolutionary processes of harmless strains of pathogens versus occasional accidents of pathogenic mutants [1]. Only explicitly stochastic models, of which the classical ODE models are mean field versions, can capture the fluctuations observed in time series data [2].

More recently it has been demonstrated that the interaction of various strains on the infection of the host with eventual cross-immunities or other interactions between host immune system and multiple strains can generate complicated dynamic attractors. A prime example is dengue fever. A first infection is often mild or even asymptomatic and leads to life long immunity against this strain. However, a subsequent infection with another strain of the virus often causes clinical complications up to life threatening conditions and hospitalization, due to ADE. More on the biology of dengue and its consequences for the detailed epidemiological model structure can be found in Aguiar and Stollenwerk [3] including literature on previous modelling attempts, see also [4]. On the biological evidence for ADE see e.g. [5]. Besides the difference in the force of infection between primary and secondary infection, parametrized by a so called ADE parameter ϕ, which has been demonstrated to show chaotic attractors in a certain parameter region, another effect, the temporary cross-immunity after a first infection against all dengue virus strains, parametrized by the temporary cross-immunity rate α, shows bifurcations up to chaotic attractors in a much wider and biologically more realistic parameter region. The model presented in the Appendix has been described in detail in [3] and has recently been analysed for a parameter value of $\alpha = 2\ year^{-1}$ corresponding to on average half a year of temporary cross immunity which is biologically plausible [6]. For increasing ADE parameter ϕ first an equilibrium which bifurcates via a Hopf bifurcation into a stable limit cycle and then after further continuation the limit cycle becomes unstable in a torus bifurcation. This torus bifurcation can be located using numerical bifurcation software based on continuation methods tracking known equilibria or limit cycles up to bifurcation points [7]. The continuation techniques and the theory behind it are described e.g. in Kuznetsov [8]. Complementary methods like Lyapunov exponent spectra can also characterize chaotic attractor [9, 10], and led ultimately to the detection of coexisting attractors to the main limit cycles and tori originated from the analytically accessible fixed point for small ϕ. Such coexisting structures are often missed in bifurcation analysis of higher dimensional dynamical systems but are demonstrated to be crucial at times in understanding qualitatively the real world data, as for example demonstrated previously in a childhood disease study [11]. In such a study first the understanding of the deterministic system's attractor structure is needed, and then eventually the interplay between attractors mediated by population noise in the stochastic version of the system gives the full understanding of the data. Here we present for the first time extended results of the bifurcation structure for various parameter values of the temporary cross immunity α in the region of biological relevance and multi-parameter bifurcation analysis. This reveals besides the torus bifurcation route to chaos also the classical Feigenbaum period doubling sequence and the origin of so called isola solutions. The symmetry of the different strains leads to symmerty breaking bifurcations of limit cycles, which are rarely described in the epidemiological literature but well known in the biochemical literature, e.g for coupled identical cells. The interplay between different numerical procedures and basic analytic insight in terms of symmetries help to understand the attractor structure of multi-strain interactions in the present case of dengue fever, and will contribute to the final understanding of dengue epidemiology including the observed fluctuations in real world data. In the literature the multi-strain interaction leading to deterministic chaos via ADE has been described previously, e.g. [12, 13] but neglecting temporary cross immunity and hence getting stuck in rather unbiological parameter regions, whereas more recently the first considerations of temporary cross immunity in rather complicated and up to now not in detail analysed models including all kinds of interations have appeared [14, 15], in this case failing to investigate closer the possible dynamical structures.

2 Dynamical system

The multistrain model under investigation can be given as an ODE system

$$\frac{d}{dt}\,\underline{x} = \underline{f}(\underline{x}, \underline{a}) \tag{1}$$

for the state vector of the epidemiological host classes $\underline{x} := (S, I_1, I_2, ..., R)^{tr}$ and besides other fixed parameters which are biologically undisputed the parameter vector of varied parameters $\underline{a} = (\alpha, \phi)^{tr}$. For a detailed description of the biological content of state variables and parameters see [3]. The ODE equations and fixed parameter values are given in the appendix. The equilibrium values $\underline{x}^*$ are given by the equilibrium condition $\underline{f}(\underline{x}^*, \underline{a}) = 0$, respectively for limit cycles $\underline{x}^*(t+T) = \underline{x}^*(t)$ with period T. For chaotic attractors the trajectory of the dynamical system reaches in the time limit of infinity the attractor trajectory $\underline{x}^*(t)$, equally for tori with irrational winding ratios. In all cases the stability can be analysed considering small perturbations $\Delta\underline{x}(t)$ around the attractor trajectories

$$\frac{d}{dt}\Delta\underline{x} = \left.\frac{d\underline{\underline{f}}}{d\underline{x}}\right|_{\underline{x}^*(t)} \cdot \Delta\underline{x} \quad . \tag{2}$$

Here, any attractor is notified by $\underline{x}^*(t)$, be it an equilibrium, periodic orbit or chaotic attractor. In this ODE system the linearized dynamics is given with the Jacobian matrix $\frac{d\underline{f}}{d\underline{x}}$ of the ODE system Eq. (1) evaluated at the trajectory points $\underline{x}^*(t)$ given in notation of $(d\underline{f}/d\underline{x})\big|_{\underline{x}^*(t)}$. The Jacobian matrix is analyzed for equilibria in terms of eigenvalues to determine stability and the loss of it at bifurcation points, negative real part indicating stability. For the stability and loss of it for limit cylces Floquet multipliers are more common (essentially the exponentials of eigenvalues), multipliers inside the unit circle indicating stability, and where they leave eventually the unit circle determining the type of limit cycle bifurcations. And for chaotic systems Lyapunov exponents are determined from the Jacobian around the trajectory, positive largest exponents showing deterministic chaos, zero largest showing limit cycles including tori, largest smaller zero indicating fixed points.

2.1 Symmetries

To investigate the bifurcation structure of the system under investigation we first observe the symmetries due to the multi-strain structure of the model. This becomes important for the time being for equilibria[1] and limit cycles. We introduce the following notation: With a symmetry transformation matrix $\mathbf{S}$

$$\mathbf{S} := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

we have the following symmetry:

$$\text{If } \underline{x}^* = \begin{pmatrix} S^* \\ I_1^* \\ I_2^* \\ R_1^* \\ R_2^* \\ S_1^* \\ S_2^* \\ I_{12}^* \\ I_{21}^* \\ R^* \end{pmatrix} \text{ is equilibrium or limit cycle, then also } \mathbf{S}\underline{x}^* = \begin{pmatrix} S^* \\ I_2^* \\ I_1^* \\ R_2^* \\ R_1^* \\ S_2^* \\ S_1^* \\ I_{21}^* \\ I_{12}^* \\ R^* \end{pmatrix} . \tag{4}$$

with $\underline{x}^*$ equilibrium values or $\underline{x}^* = \underline{x}^*(t)$ limit cycle for all times $t \in [0, T]$. For the right hand side $\underline{f}$ of the ODE system Eq. (1) the kind of symmetry found above is

[1] Equilibria are often called fixed points in dynamical systems theory, here we try to avoid this term, since in symmetry the term *fixed* is used in a more specific way, see below.

called $\mathbb{Z}_2$-symmetry when the following equivariance condition holds

$$\underline{f}(\mathbf{S}\underline{x},\underline{a}) = \mathbf{S}\underline{f}(\underline{x},\underline{a}) \tag{5}$$

with $\mathbf{S}$ a matrix that obeys $\mathbf{S} \neq \mathbf{I}$ and $\mathbf{S}^2 = \mathbf{I}$, where $\mathbf{I}$ is the unit matrix. Observe that besides $\mathbf{S}$ also $\mathbf{I}$ satisfies (5). The symmetry transformation matrix $\mathbf{S}$ in Eq. (3) fulfills these requirements. It is easy to verify that the $\mathbb{Z}_2$-equivariance conditions Eq. (5) and the properties of $\mathbf{S}$ are satisfied for our ODE system. In Seydel [16] a simplified version of the famous Brusselator that shows this type of symmetry is discussed. There, an equilibrium and also a limit cycle show a pitchfork bifurcation with symmetry breaking.

An equilibrium $\underline{x}^*$ is called *fixed* when $\mathbf{S}\underline{x}^* = \underline{x}^*$ (see [8]). Two equilibria $\underline{x}^*, \underline{y}^*$ where $\mathbf{S}\underline{x}^* \neq \underline{x}^*$, are called $\mathbf{S}$-conjugate if their corresponding solutions satisfy $\underline{y}^* = \mathbf{S}\underline{x}^*$ (and because $\mathbf{S}^2 = \mathbf{I}$ also $\underline{x}^* = \mathbf{S}\underline{y}^*$). For limit cycles a similar terminology is introduced. A periodic solution is called *fixed* when $\mathbf{S}\underline{x}^*(t) = \underline{x}^*(t)$ and the associated limit cycles are also called *fixed* [8]. There is another type of periodic solution that is not fixed but called *symmetric* when

$$\mathbf{S}\underline{x}^*(t) = \underline{x}^*\left(t + \frac{T}{2}\right) \tag{6}$$

where T is the period. Again the associated limit cycles are also called *symmetric*. Both types of limit cycles L are $\mathbf{S}$-invariant as curves : $\mathbf{S}L = L$. That is, in the phase-plane where time parameterizes the orbit, the cycle and the transformed cycle are equal. A $\mathbf{S}$-invariant cycle is either fixed or symmetric. Two noninvariant limit cycles ($\mathbf{S}L \neq L$) are called $\mathbf{S}$-conjugate if their corresponding periodic solutions satisfy $\underline{y}^*(t) = \mathbf{S}\underline{x}^*(t)$, $\forall t \in \mathbb{R}$. The properties of the symmetric systems and the introduced terminology are used below with the interpretation of the numerical bifurcation analysis results. We refer to [8] for an overview of the possible bifurcations of equilibria and limit cycles of $\mathbb{Z}_2$-equivariant systems.

3 Bifurcation diagrams for various α values

We show the results of the bifurcation analysis in bifurcation diagrams for several α values, varying ϕ continuously. Besides the previously investigated case of $\alpha = 2\ year^{-1}$, we show also a case of smaller and a case of larger α value, obtaining more information on the bifurcations possible in the model as a whole. The above mentioned symmetries help in understanding the present bifurcation structure.

3.1 Bifurcation diagram for $\alpha = 3$

For $\alpha = 3$ the one-parameter bifurcation diagram is shown in Fig. 1 a). Starting with $\phi = 0$ there is a stable fixed equilibrium, fixed in the above mentioned notion for symmetric systems. This equilibrium becomes unstable at a Hopf bifurcation H at $\phi = 0.164454$. A stable fixed limit cycle originates at this Hopf bifurcation. This limit cycle shows a supercritical pitch-fork bifurcation P^-, i.e. a bifurcation of a limit cycle with Floquet multiplier 1, splitting the original limit cycle into two new ones. Besides the now unstable branch two new branches originate for the pair of conjugated limit cycles. The branches merge again at another supercritical pitch-fork bifurcation P^-, after which the limit cycle is stable again for higher ϕ-values. The pair of $\mathbf{S}$-conjugate limit cycles become unstable at a torus bifurcation TR at $\phi = 0.89539$.

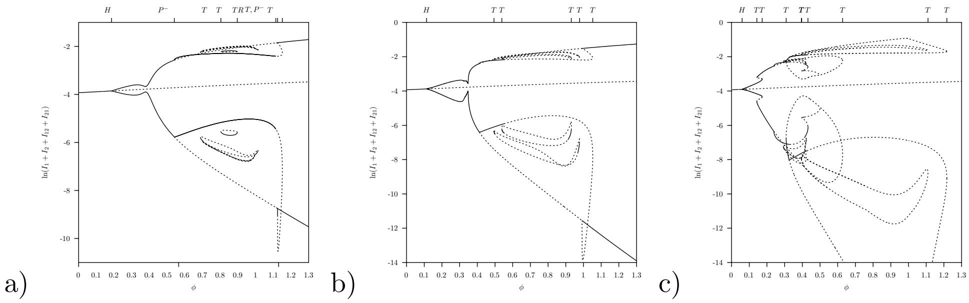


Figure 1: a) $\alpha = 3$: Equilibria or extremum values for limit cycles for logarithm of total infected $I_1 + I_2 + I_{12} + I_{21}$. Solid lines denote stable equilibria or limit cycles, dashed lines unstable equilibria or periodic-one limit cycles. Hopf bifurcation H around $\phi = 0.16$ two pitchfork bifurcations P^- and a torus bifurcation TR. Besides this main bifurcation structure we found coexisting tangent bifurcations T between which some of the isolas live, see especially the one between $\phi = 0.71$ and 0.79. Additionally found flip bifurcations are not marked here, see text. b) $\alpha = 2$: In this case we have a Hopf bifurcation H at $\phi = 0.11$, and besides the similar structure as found in a) also more separated tangent bifurcations T at $\phi = 0.494$, 0.539, 0.931, 0.978 and 1.052 c) $\alpha = 1$: Here we have the Hopf bifurcation at $\phi = 0.0598$ and thereafter many tangent bifurcations T, again with coexisting limit cylces.

Besides this main bifurcation pattern we found two isolas, that is an isolated solution branch of limit cycles [17]. These isola cycles L are not $\mathbf{S}$-invariant, that is $\mathbf{S}L \neq L$. Isolas consisting of isolated limit cycles exist between two tangent bifurcations. One isola consists of a stable and an unstable branch. The other shows more complex bifurcation patterns. There is no full stable branch. For $\phi = 0.60809$ at the tangent bifurcation T a stable and an unstable limit cycle collide. The stable branch becomes unstable via a flip bifurcation or periodic doubling bifurcation F, with Floquet multiplier (-1), at $\phi = 0.61918$ which is also pitchfork bifurcation for the period-two limit cycles. At the other end of that branch at the tangent bifurcation T at $\phi = 0.89768$ both colliding limit cycles are unstable. Close to this point at one branch there is a torus bifurcation TR, also called Neimark-Sacker bifurcation, at $\phi = 0.89539$ and a flip bifurcation F at $\phi = 0.87897$ which is again a pitchfork bifurcation P for the period-two limit cycles. Contiuation of the stable branch originating for the flip bifurcation F at $\phi = 0.61918$ gives another flip bifurcation F at $\phi = 0.62070$ and one closed to the other end at $\phi = 0.87897$, namely at $\phi = 0.87734$. These results suggest that for this isola two classical routes to chaos can exist, namely via the torus or Neimark-Sacker bifurcation where the dynamics on the originating torus is chaotic, and the cascade of period doubling route to chaos.

3.2 Bifurcation diagram for $\alpha = 2$

For $\alpha = 2$ the one-parameter bifurcation diagram is shown in Fig. 1 b). The stable fixed equilibrium becomes unstable at a supercritical Hopf bifurcation H at $\phi = 0.1132861$ where a stable fixed limit cycle originates. This stable limit cycle becomes unstable at a superciritcal pitchfork bifurcation point P^- at $\phi = 0.4114478$ for a limit cycle. This point marks the origin of a pair of **S**-conjugate stable limit cycles besides the now unstable fixed limit cycle. Here one has to consider the two infected subpopulations I_1 and I_2 to distinguish the conjugate limit cycles. Because the two variables I_1 and I_2 are interchangeable this can also be interpreted as the stable limit cycles for the single variable say I_1. The fixed stable equilibrium below the Hopf bifurcation where we have $I_1^* = I_2^*$, $R_1^* = R_2^*$, $S_1^* = S_2^*$ and $I_{12}^* = I_{21}^*$ is a fixed equilibrium. For the fixed limit cycle in the parameter interval between the Hopf bifurcation and the pitchfork bifurcation we have $I_1^*(t) = I_2^*(t)$, $R_1^*(t) = R_2^*(t)$, $S_1^*(t) = S_2^*(t)$ and $I_{12}^*(t) = I_{21}^*(t)$. This means that at the Hopf bifurcation H the stable fixed equilibrium becomes an unstable fixed equilibrium. In the parameter interval between the two pitchfork bifurcations P^- at $\phi = 0.4114478$ and subcritical P^+ at $\phi = 0.9921416$, two stable limit cycles coexist and these limit cycles are **S**-conjugate. At the pitchfork bifurcation points the fixed limit cycle becomes unstable and remains fixed, and two stable **S**-conjugate limit cycles originate (see [8, Theorem 7.7]). The invariant plane $I_1 = I_2, R_1 = R_2, S_1 = S_2, I_{12} = I_{21}$ forms the separatrix between the pair of stable **S**-conjugate limit cycles $x^*(t)$ and $\mathbf{S}x^*(t)$, $\forall t \in \mathbb{R}$. The initial values of the two state variables $S(t_0)$ and $R(t_0)$ together with the point on the invariant plane, determine to which limit cycle the system converges. Continuation of the stable symmetric limit cycle gives a torus or Neimark-Sacker bifurcation at point denoted by TR at $\phi = 0.5506880$. At his point the limit cycles become unstable because a pair of complex-conjugate multipliers crosses the unit circle. Observe that at this point in the time series plot [3, there Fig. 12] the chaotic region starts. In [18] the following route to chaos, namely the sequence of Neimark-Sacker bifurcations into chaos, is mentioned. Increasing the bifurcation parameter ϕ along the now unstable pair of **S**-conjugate limit cycles leads to a tangent bifurcation T at $\phi = 1.052418$ where a pair of two unstable limit cycles collide. This branch terminates at the second pitchfork bifurcation point denoted by P^+ at $\phi = 0.9921416$. Because the first fold point gave rise to a stable limit cycle and this fold point to an unstable limit cycle we call the first pitchfork bifurcation supercritical and the latter pitchfork bifurcation subcritical. These results agree very well with the simulation results shown in the bifurcation diagram for the maxima and minima of the overall infected [3, there Fig. 15]. Notice that AUTO [7] calculates only the global extrema during a cycle, not the local extrema. Fig. 1 b) shows also two isolas similar to those for $\alpha = 3$ in Fig. 1 a).

3.3 Bifurcation diagram for $\alpha = 1$

For $\alpha = 1$ the bifurcation diagram is shown in Fig 1 c). In the lower ϕ parameter range there is bistability of two limit cycles in an interval bounded by two tangent bifurcations

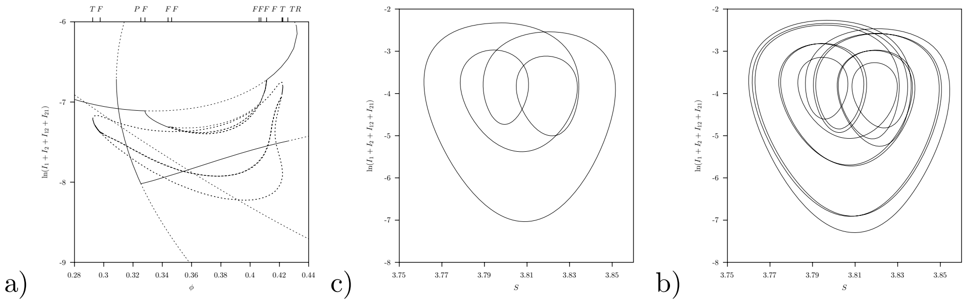


Figure 2: a) $\alpha = 1$. Detail of Fig. 1 c). We find pitchfork bifurcations P at $\phi = 0.239$ and 0.325, flip bifurcations F at $\phi = 0.298$, 0.328,0.344,0.346, 0.406, 0.407, 0.411 and 0.422, further tangent bifurcations T at $\phi = 0.292$, 0.346 and 0.422. Four almost coexisting bifurcations, namely F's at $\phi = 0.4112590$. b) and c) state space-plots of susceptibles and logarithm of infected for $\alpha = 1$ and $\phi = 0.294$ for two coexisting stable limit cycles.

T. The stable manifold of the intermediate saddle limit cycle acts as a separatrix. Inceasing ϕ the stable limit cycles become unstable at the pitchfork bifurcation P at $\phi = 0.2390695$. Following the unstable primary branch, for larger values of ϕ we observe an open loop bounded by two tangent bifurcations T. The extreme value for ϕ is at $\phi = 0.6279042$. Then lowering ϕ there is a pitchfork bifurcation P at $\phi = 0.5016112$. Later we will return to the description of this point. Lowering ϕ further the limit cycle becomes stable again at the tangent bifurcations T at $\phi = 0.3086299$. Increasing ϕ this limit cycle becomes unstable again at the pitchfork bifurcation P at $\phi = 0.3253242$.

Continuation of the secondary branch of the two **S**-conjugated limit cycles from this point reveals that the stable limit cycle becomes unstable at a torus bifurcation TR at $\phi = 0.4257346$. The simulation results depicted in [3, Fig. 13] show that there is chaos beyond this point. The secondary pair of **S**-conjugate limit cycles that originate from pitchfork bifurcation P at $\phi = 0.2390695$ becomes unstable at a flip bifurcation F. Increasing ϕ further it becomes stable again at a flip bifurcation F. Below we return to the interval between these two flip bifurcations. The stable part becomes unstable at a tangent bifurcation T, then continuing, after a tangent bifurcation T and a Neimark-Sacker bifurcation TR. This bifurcation can lead to a sequence of Neimark-Sacker bifurcations into chaos. The unstable limit cycles terminates via a tangent bifurcation F where the primary limit cycle possesses a pitchfork bifurcation P at $\phi = 0.5016112$. At the flip bifurcation F the cycle becomes unstable and a new stable limit cycle with double period emanates. The stable branch becomes unstable at a flip bifurcation again. We conclude that there is a cascade of period doubling route to chaos. Similarly this happens in reversed order ending at the flip bifurcation where the secondary branch becomes stable again.

Fig. 2 a) gives the results for the interval $0.28 \leq \phi \leq 0.44$ where only the minima are show. In this plot also a "period three" limit cycle is shown. In a small region it is stable and coexists together with the "period one" limit cycle. The cycles are shown in Fig. 2 b) and c) for $\phi = 0.294$. The one in c) looks like a period-3 limit cycle. In

Fig. 2 continuation of the limit cycle gives a closed graph bounded at the two ends by trangent bifurcations T where a stable and an unstable limit cycle collide. The intervals where the limit cycle is stable, are on the other end bounded by flip bifurcations F. One unstable part intersects the higher period cycles that originate via the cascade of period doubling between the period-1 limit cycle flip bifurcations F at $\phi = 0.3281636$ and $\phi = 0.4112590$. This suggest that the period-3 limit cycle is associated with a "period-3 window" of the chaotic attractor. We conjecture that this interval is bounded by two homoclinic bifurcations for a period-3 limit cycle (see [19, 20, 21, 22]). The bifurcation diagram shown in [3, there Fig. 13] shows the point where the chaotic attractor disappears abruptly, possible at one of the two homoclinic bifurcations. In that region the two conjugated limit cycles that originate at the pitchfork bifurcation P at $\phi = 0.3253242$ are the attractors. These results suggest that there are chaotic attractors associated with the period-1 limit cycle, one occurs via a cascade of flip bifurcations originating from the two ends at $\phi = 0.3281636$ and $\phi = 0.4112590$ and one via a Neimark-Sacker bifurcation TR at $\phi = 0.4257346$.

4 Two-parameter diagram

We will now link the three studies of the different α values by investigating a two-parameter diagram for ϕ and α, concentrating especially on the creation of isolated limit cycles, which sometimes lead to further bifurcations inside the isola region. Fig. 3 gives a two-parameter bifurcation diagram where ϕ and α are the free parameters. For low ϕ-values there is the Hopf bifurcation H and all other curves are tangent bifurcation curves.

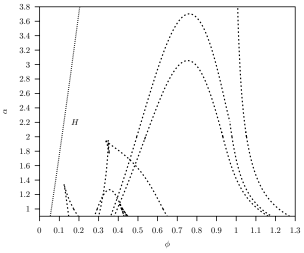


Figure 3: Two-dimensional parameter bifurcation diagram with ϕ and α as parameters. Only one Hopf bifurcation (dotted lines) and many tangent bifurcation curves (dashed lines) are shown in the range $\alpha \in [1, 4]$. The isolated limit cycles originate above $\alpha = 3$. For lower values of α periodic doubling routes to chaos originate.

Isolas appear or disappears upon crossing an isola variety. At an elliptic isola point an isolated solution branch is born, while at a hyperbolic isola point an isolated solution branch vanishes by coalescence with another branch [17]. From Fig. 3 we see that at two values of $\alpha > 3$ isolas are born. Furthermore, period doubling bifurcations appear for lower α values, indicating the Feigenbaum route to chaos. However, only the calculation of Lyapunov exponents, which are discussed in the next section, can clearly indicate chaos.

5 Lyapunov spectra for various α values

The Lyapunov exponents are the logarithms of the eigenvalues of the Jacobian matrix along the integrated trajectories, Eq. (2), in the limit of large integration times. Besides for very simple iterated maps no analytic expressions for chaotic systems can be given for the Lyapunov exponents. For the calculation of the iterated Jacobian matrix and its eigenvalues, we use the QR decomposition algorithm [23].

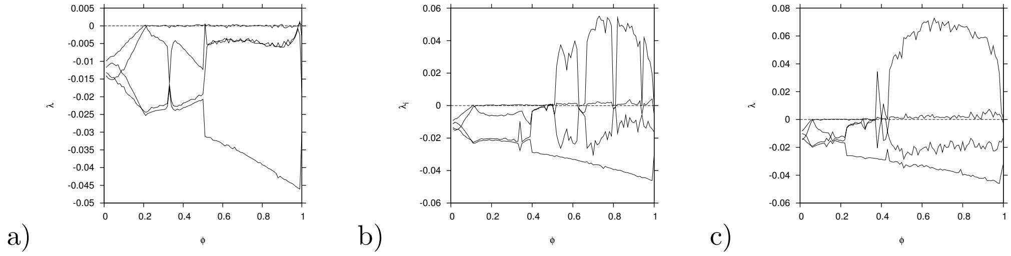


Figure 4: *Spectrum of the four largest Lyapunov exponents with changing parameter ϕ and a) fixed $\alpha = 4$, b) $\alpha = 2$ and c) $\alpha = 1$.*

In Fig. 4 we show for various α values the four largest Lyapunov exponents in the ϕ range between zero and one. For $\alpha = 4$ in Fig. 4 a) we see for small ϕ values fixed point behaviour indicated by a negative largest Lyapunov exponent up to around $\phi = 0.2$. There, at the Hopf bifurcation point, the largest Lyapunov exponent becomes zero, indicating limit cycle behaviour for the whole range of ϕ, apart from the final bit before $\phi = 1$, where a small spike with positive Lyapunov exponent might be present, but difficult to distinguish from the noisy numerical background.

For $\alpha = 2$ in Fig. 4 b) however, we see a large window with positive largest Lyapunov exponent, well separated from the second largest being zero. This is s clear sign of deterministically chaotic attractors present for this ϕ range. Just a few windows with periodic attractors, indicated by the zero largest Lyapunov exponent are visible in the region of $0.5 < \phi < 1$. For smaller ϕ values we observe qualitatively the same behaviour as already seen for $\alpha = 4$. For the smaller value of $\alpha = 1$ in Fig. 4 c) the chaotic window is even larger than for $\alpha = 2$. Hence deterministic chaos is present for temporary cross immunity in the range around $\alpha = 2$ $year^{-1}$ in the range of ϕ between zero and one.

6 Conclusions

We have presented a detailed bifurcation analysis for a multi-strain dengue fever model in terms of the ADE parameter ϕ, in the previously not well investigated region between zero and one, and a parameter for the temporary cross immunity α. The symmetries implied by the strain structure, are taken into account in the analysis. Many of the possible bifurcations of equilibria and limit cycles of $\mathbb{Z}_2$-equivariant systems can be distinguished. Using AUTO [7] the different dynamical structures were calculated.

Future time series analysis of epidemiological data has good chances to give insight into the relevant parameter values purely on topological information of the dynamics, rather than classical parameter estimation of which application is in general restricted to farely simple dynamical scenarios.

Acknowledgements

This work has been supported by the European Union under the Marie Curie grant MEXT-CT-2004-14338. We thank Gabriela Gomes and Luis Sanchez, Lisbon, for scientific support.

References

[1] N. Stollenwerk and V.A.A. Jansen, *Evolution towards criticality in an epidemiological model for meningococcal disease*, Physics Letters A **317** (2003) 87–96.

[2] N. Stollenwerk, M.C.J. Maiden and V.A.A. Jansen, *Diversity in pathogenicity can cause outbreaks of menigococcal disease*, Proc. Natl. Acad. Sci. USA **101** (2004) 10229–10234.

[3] M. Aguiar and N. Stollenwerk, *A new chaotic attractor in a basic multi-strain epidemiological model with temporary cross-immunity*, arXiv:0704.3174v1 [nlin.CD] (2007) (accessible electronically at http://arxive.org).

[4] E. Massad, M. Chen, S. Ma, C. J. Struchiner, N. Stollenwerk, and M. Aguiar, *Scale-free Network for a Dengue Epidemic*, Applied Mathematics and Computation **159** (2008) 376–381.

[5] S.B. Halstead, *Neutralization and antibody-dependent enhancement of dengue viruses*, Advances in Virus Research **60** (2003) 421–67.

[6] M. Aguiar, B.W. Kooi, and N. Stollenwerk, *Epidemiology of dengue fever: A model with temporary cross-immunity and possible secondary infection shows bifurcations and chaotic behaviour in wide parameter regions*, submitted (2008).

[7] E.J. Doedel, R.C. Paffenroth, A.R. Champneys, T.F. Fairgrieve, Y.A. Kusnetsov, B. Sandstede, B. Oldeman, X.J. Wang, and C. Zhang, *AUTO 07P – Continuation and bifurcation software for ordinary differential equations*, Technical Report: Concordia University, Montreal, Canada (2007) (accessible electronically at http://indy.cs.concordia.ca/auto/).

[8] Y.A. Kuznetsov, *Elements of Applied Bifurcation Theory* Applied Mathematical Sciences **112**, Springer-Verlag, 3 edition, New York, 2004.

[9] D. Ruelle,*Chaotic Evolution and Strange Attractors*, Cambridge University Press, Cambridge, 1989.

[10] E. Ott, *Chaos in Dynamical Systems*, Cambridge University Press, Cambridge, 2002.

[11] F.R. Drepper, R. Engbert and N. Stollenwerk, *Nonlinear time series analysis of empirical population dynamics*, Ecological Modelling **75/76** (1994) 171–181.

[12] N. Ferguson, R. Anderson, and S. Grupta, *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.

[13] L. Billings, B.I. Schwartz, B.L. Shaw, M. McCrary, D.S. Burke and T.A.D. Cummings, *Instabilities in multiserotype disease models with antibody-dependent enhancement*, Journal of Theoretical Biology **246** (2007) 18–27.

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

[15] Y. Nagao and K. Koelle, *Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever*, Proc. Natl. Acad. Sci **105** (2008) 2238–2243.

[16] R. Seydel, *Practical bifurcation and stability analysis-from equilibrium to chaos*, Springer-Verlag, New York, 1994.

[17] M. Golubitsky and D.G. Schaeffer, *Singularities and groups in bifurcation theory*, Springer, New York, 1985.

[18] D. Albers and J. Sprott, *Routes to chaos in high-dimensional dynamical systems: A qualitative numerical study*, Physica D **223** (2006) 194–207.

[19] M.P. Boer,B.W. Kooi and S.A.L.M. Kooijman, *Homoclinic and heteroclinic orbits in a tri-trophic food chain*, Journal of Mathematical Biology **39** (1999) 19–38.

[20] M.P. Boer, B.W. Kooi and S.A.L.M. Kooijman, *Multiple attractors and boundary crises in a tri-trophic food chain*, Mathematical Biosciences **169** (2001) 109–128.

[21] B.W. Kooi and M.P. Boer, *Chaotic behaviour of a predator-prey system*, Dynamics of Continuous, Discrete and Impulsive Systems, Series B: Applications and Algorithms **10** (2003) 259–272.

[22] B.W. Kooi, L.D.J. Kuijper and S.A.L.M. Kooijman, *Consequence of symbiosis for food web dynamics*, Journal of Mathematical Biology **49** (2004) 227–271.

[23] J.P. Eckmann, S. Oliffson-Kamphorst, D. Ruelle and S. Ciliberto, *Liapunov exponents from time series*, Phys. Rev. A **34** (1986) 4971–9.

7 Appendix: Epidemic model equations

The complete system of ordinary differential equations for a two strain epidemiological system allowing for differences in primary versus secondary infection and temporary cross immunity is given by

$$\begin{aligned}
\frac{d}{dt}S &= -\frac{\beta}{N}S(I_1+\phi I_{21})-\frac{\beta}{N}S(I_2+\phi I_{12})+\mu(N-S)\\
\frac{d}{dt}I_1 &= \frac{\beta}{N}S(I_1+\phi I_{21})-(\gamma+\mu)I_1\\
\frac{d}{dt}I_2 &= \frac{\beta}{N}S(I_2+\phi I_{12})-(\gamma+\mu)I_2\\
\frac{d}{dt}R_1 &= \gamma I_1-(\alpha+\mu)R_1\\
\frac{d}{dt}R_2 &= \gamma I_2-(\alpha+\mu)R_2\\
\frac{d}{dt}S_1 &= -\frac{\beta}{N}S_1(I_2+\phi I_{12})+\alpha R_1-\mu S_1\\
\frac{d}{dt}S_2 &= -\frac{\beta}{N}S_2(I_1+\phi I_{21})+\alpha R_2-\mu S_2\\
\frac{d}{dt}I_{12} &= \frac{\beta}{N}S_1(I_2+\phi I_{12})-(\gamma+\mu)I_{12}\\
\frac{d}{dt}I_{21} &= \frac{\beta}{N}S_2(I_1+\phi I_{21})-(\gamma+\mu)I_{21}\\
\frac{d}{dt}R &= \gamma(I_{12}+I_{21})-\mu R \quad .
\end{aligned} \tag{7}$$

For two different strains, 1 and 2, we label the SIR classes for the hosts that have seen the individual strains. Susceptibles to both strains (S) get infected with strain 1 (I_1) or strain 2 (I_2), with infection rate β. They recover from infection with strain 1 (becoming temporary cross-immune R_1) or from strain 2 (becoming R_2), with recovery rate γ etc.. With rate α, the R_1 and R_2 enter again in the susceptible classes (S_1 being immune against strain 1 but susceptible to 2, respectively S_2), where the index represents the first infection strain. Now, S_1 can be reinfected with strain 2 (becoming I_{12}), meeting I_2 with infection rate β or meeting I_{12} with infection rate $\phi\beta$, secondary infected contributing differently to the force of infection than primary infected, etc..

We include demography of the host population denoting the birth and death rate by μ. For constant population size N we have for the immune to all strains $R = N-(S+I_1+I_2+R_1+R_2+S_1+S_2+I_{12}+I_{21})$ and therefore we only need to consider the first 9 equations of Eq. (7), giving 9 Lyapunov exponents. In our numerical studies we take the population size equal to $N = 100$ so that numbers of susceptibles, infected etc. are given in percentage. As fixed parameter values we take $\mu = (1/65)\ year^{-1}$, $\gamma = 52\ year^{-1}$, $\beta = 2\cdot\gamma$. The parameters ϕ and α are varied.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008

A Meshless Approach for Electromagnetic Simulation of Metallic Carbon Nanotubes

G. Ala[1], E. Francomano[2] A.Spagnuolo[1], A. Tortorici[2]
[1] *Dipartimento di Ingegneria Elettrica, Elettronica e delle Telecomunicazioni, Università degli Studi di Palermo, Italy*
[2] *Dipartimento di Ingegneria Informatica, Università degli Studi di Palermo, Italy*

emails: guido.ala@unipa.it, e.francomano@unipa.it

Abstract

In this paper, a study on the electromagnetic behaviour of a single wall carbon nanotube model is described. The electrons available for conduction are treated as a thin cylindrical layer fluid and their motion is described by means of classical hydrodynamics equations in linearized form. These equations are solved in time domain using the Smoothed Particle Hydrodynamics method. The method, suitably handled runs on GRID environment.

Key words: carbon nanotubes, electromagnetics, meshless methods, Smoothed Particle Hydrodynamic, GRID computing.
MSC2000: 65C20, 76N20

1. Introduction

The technological evolution in electronics has become more and more rapid, shrinking the dimension of chips and integrated circuits (IC) and raising the upper frequency limit. In such devices, the traditional conductor materials (copper and noble metals) are no more suitable, due to electromagnetic (EM) phenomena which do not fit with the correct working. The effects of eddy currents and the inducted high-frequency EM fields are the hardest obstacles which limit the development of smaller ICs.

Today, chip producers are oriented in employing alternative materials for the next generation products. One of the most promising materials is carbon, in form of nanotubes.

Singular Wall Nanotubes (SWNT) and Multi Wall Nanotubes (MWNT) seem to have an optimal behaviour both electromagnetically and thermically. Besides, they have great resilience and low cost when the industrial process for large scale production is optimized,.

In the last years, field effect transistor (FET), dipole antennas and interconnects realized by carbon nanotubes have been presented and discussed. The nanotubes have high electrical and thermal conductivity, bear impressive current densities and have good mechanical properties[1],[2].

In this paper, the electromagnetic behaviour of carbon nanotubes is investigated assuming the π-electron motion governed by classical equations of dynamics [3],[4]. The problem is described by partial differential equations in time domain. The numerical solution is performed by revisiting the Smoothed Particle Hydrodynamics (SPH) method generally used in fluid dynamics context [5]. The proposed formulation is applied to determine the current flowing in a carbon nanotube for different frequencies of the voltage source.

The GRID environment has shown to be suitable to minimize the computational runtime due to the high number of frequencies under investigation.

2. Structure of Single Wall Nanotubes

Carbon nanotubes are obtained rolling bi-dimensional sheets of graphene (an one-atom-thick graphite sheet) along a generic axis [1],[2].

In order to effectively describe the nanotube structure, let us consider a graphene sheet (fig.1).

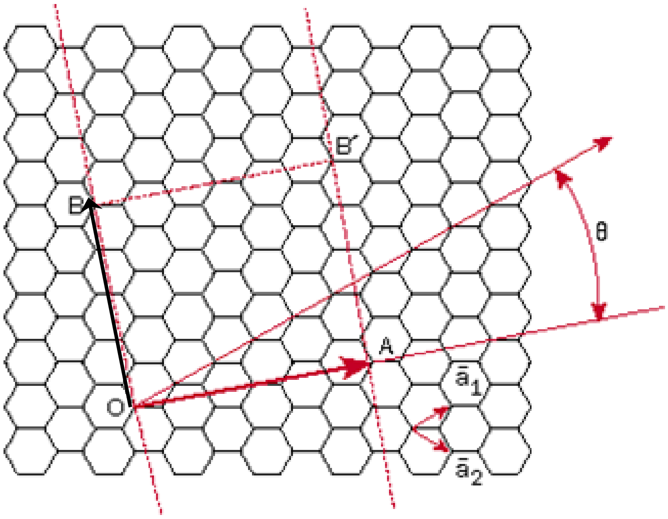


Fig.1 - Graphene sheet. Vectors and geometrical parameters.

The nanotube is described by the vectors $\vec{a}_1$ and $\vec{a}_2$, with $|\vec{a}_1| = |\vec{a}_2| = a_0 = 0.246$ nm, forming a 60° angle. The graphene sheet is then rolled with

OA = $n_1\vec{a}_1 + n_2\vec{a}_2$ being the circumference of the tube. This vector, called *chiral vector*, is often described by the pair *(n_1, n_2)* and denotes univocally that nanotube.

In fig. 1, **OA** = $4\vec{a}_1 + 2\vec{a}_2$ and the resulting nanotube will be denoted as *(4,2)* nanotube.

The *translational vector* **OB**, perpendicular to **OA**, represents the distance after which the surface pattern repeats itself. The rectangle **OABB'** is called *unit cell*.

The properties of the nanotube depend on the angle between the chiral vector and $\vec{a}_1$ (*chiral angle*). Nanotubes with similar diameter but different chiral vectors show totally different behaviours.

In particular, *(n,0)* nanotubes are called *zig-zag* nanotubes and show a semi-conductor behaviour, whereas *(n,n)* nanotube are called *armchair* nanotubes and behave like metallic conductors.

The chiral angle can be expressed as:

$$\theta = \arccos\frac{\vec{a}_1 \cdot \vec{c}}{|\vec{a}_1| \cdot \vec{c}} = \frac{n_1 + n_2/2}{\sqrt{n_1^2 + n_1 n_2 + n_2^2}}, \qquad (1)$$

in which $\vec{c} = |\mathbf{OA}| = a_0\sqrt{n_1^2 + n_1 n_2 + n_2^2}$ **.**

This angle is 0° for zig-zag nanotubes and 30° for armchair ones. Both the chiral vector and the chiral angle determine the geometrical parameters of the tube (diameter, number of carbon atoms in the unit cell, length of the translational vector, etc.) as shown in fig.2.

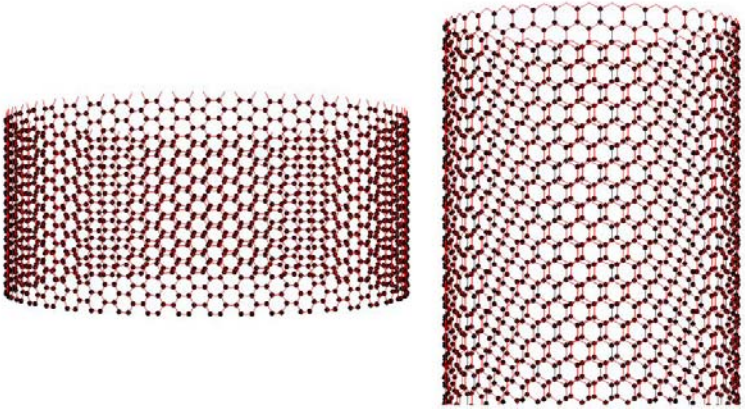


Fig.2 - A *(40,40)* nanotube on the left and a *(40,0)* nanotube on the right.

The diameter can be expressed as:

$$d = \frac{|\vec{c}|}{\pi} = \frac{a_0}{\pi}\sqrt{n_1^2 + n_1 n_2 + n_2^2} = \frac{a_0}{\pi}\sqrt{N}\,, \qquad (2)$$

with $N=n^2_1+n_1n_2+n^2_2$.
The translational vector can be expressed as follows:

$$|\mathbf{OB}|=\frac{\sqrt{3N}}{n\Re}a_0\,,\qquad(3)$$

where $\Re=3$ if $(n_1-n_2)/3n$ is an integer, otherwise, $\Re=1$ (n is the greatest common divisor of n_1 and n_2).
Each graphene cell contains 2 carbon atoms and the number of carbon atoms, n_C, in the unit cell can be expressed as the ratio between the lateral area of the nanotube S_L and the area of the singular graphene cell S_g:

$$n_C=2\frac{S_L}{S_g}=\frac{4N}{n\Re}\,.\qquad(4)$$

2. The model

Studies on the EM behaviour of carbon nanotubes by considering them as transmission lines have been carried out [6]. More recently, a model based on the hydrodynamics of an electron gas has been proposed.
Because of the structure of the carbon-carbon bond in nanotubes, each atom is bonded with other three atoms, thus leaving only an electron available for the conduction (π-electron) [7].
Therefore, the available electrons can be regarded as a single layer electron gas and treated according to the classical motion equations of fluid dynamics [3],[4].
Let us consider the classical equations of conservation of density and momentum in linearized form,

$$\frac{\partial\delta n}{\partial t}+n_0\nabla_{\parallel}\cdot\vec{v}=0\qquad(5)$$

$$n_0\frac{\partial\vec{v}}{\partial t}=-\frac{1}{m_{eff}}\nabla_{\parallel}\delta p-e\frac{n_0}{m_{eff}}\vec{e}_{\parallel}-\frac{1}{\tau}n_0\vec{v}\,,\qquad(6)$$

being n_0 the density number of electrons able to conduct at the equilibrium, $\vec{v}$ the speed of the electrons, p the pressure due to the fluid motion, e and m_{eff} the electron charge and effective mass, $\vec{e}$ is the electric field directed along the tube axis and τ is the relaxation time respectively.. The symbol $_{\parallel}$ refers to the direction of the tube axis.
By expressing the pressure as

$$\delta p=\left(\frac{\partial p}{\partial n}\right)_{n=n_0}\cdot\delta n=m_{eff}c_S^2\cdot\delta n\,,\qquad(7)$$

in which c_S is the thermodynamic speed of sound in the electron fluid, and by expressing the current and superficial charge densities on the lateral surface of the nanotube as

$$\begin{aligned} &\sigma = -e\delta n(\vec{r}_S, t) \\ &\vec{j} = -en_0\vec{v}(\vec{r}_S, t)\text{'} \end{aligned} \qquad (8)$$

in wich $\vec{r}_S$ is a vector identifying a generical point on the lateral surface of the tube, the equations (5) and (6) become:

$$\frac{\partial \sigma}{\partial t} + \nabla_{\|} \cdot \vec{j} = 0 \qquad (9)$$

and

$$\frac{\partial \vec{j}}{\partial t} + \frac{1}{\tau}\vec{j} + c_S^2 \nabla_{\|} \sigma = e^2 \frac{n_0}{m_{eff}} \vec{e}_{\|}\,. \qquad (10)$$

These equations hold only for tubes with n_1,n_2 < 50 and with axial length greater than the tube diameter.

The values for the parameters c_S, τ and $\frac{n_0}{m_{eff}}$ are detailed in [7].

3. The numerical approach

To solve both equations in time domain, the *Smoothed Particle Hydrodynamics* (SPH) method is used. The SPH method is a powerful instrument for investigating the motion of fluids and gases. It is a meshless method since it does not require a computational grid with fixed nodes[5]. The problem domain is discretized by mean of "particles" and the function values and their derivatives for each particle are evaluated by using the informations of "near" particles (fig.3).

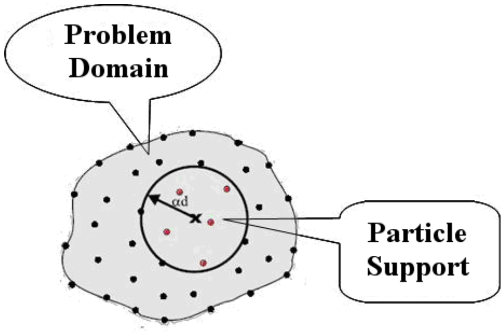


Fig.3 - Problem domain and particle support.

Namely, the integral approximation of a function *f(x)* is

$$< f(\vec{x}) >= \int_{\Omega} f(\vec{x}')W(\vec{x}-\vec{x}',h)d\vec{x}', \qquad (11)$$

where $W(\vec{x}-\vec{x}',h)$ is a suitable kernel function, defined only in the particle support, *Ω* is the problem domain, *h* is the smoothing length and $\vec{x}, \vec{x}'$ are the position vectors of the particles.

The SPH method is used to approximate the spatial derivatives of equations (9) and (10) [8]. The time derivatives are evaluated with a finite difference scheme.

In the simulation, the cubic B-Spline presented in [5] is used. The nanotube is discretized by 4000 particles, each accounting for 8 electrons scattered on the nanotube surface. Each particle is located in the center of mass of the 8-electron group.

A unitary voltage source at different frequencies is applied and the current response is measured.

The simulations at different frequencies run on GRID environment by means of the Message Passing Interface (MPI) paradigm.

Acknowledgements

This work makes use of results produced by the PI2S2 Project managed by the Consorzio COMETA, a project co-funded by the Italian Ministry of University and Research (MIUR) within the Piano Operativo Nazionale “Ricerca Scientifica, Sviluppo Tecnologico, Alta Formazione” (PON 2000-2006). More information is available at http://www.pi2s2.it and http://www.consorzio-cometa.it.

4. References

[1] R. SAITO, G. DRESSELHAUS & M.S. DRESSELHAUS, *Physical properties of carbon nanotubes*, World Scientific (1998).

[2] S. REICH, C. THOMSEN, J. MAULTZSCH, *Carbon Nanotubes*, Wiley-VCH, 2004.

[3] A. L. FETTER, *Electrodynamics of a layered electron gas. I. Single layer*, Ann. Phys., vol. 81, pages 367-393, 1973.

[4] A. L. FETTER, *Electrodynamics of a layered electron gas. II.Periodic array*, Ann. Phys., vol. 88, pages 1-25, 1974.

[5] G. R. LIU, M. B. LIU, *Smoothed Particle Hydrodynamics*, World Scientific, 2003.

[6] G. W. HANSON, *Current on an Infinitely Long Carbon Nanotube Antenna Excited by a Gap Generator*, IEEE Transactions on Antennas and Propagation, vol.54, pages 76-81, 2006.

[7] G. MIANO, F. VILLONE, *An Integral Formulation for the Electrodynamics of Metallic Carbon Nanotubes Based on a Fluid Model*, IEEE Transactions on Antennas and Propagation, vol. 54, pages 2713-2724, 2006.

[8] G. ALA, E. FRANCOMANO, A. TORTORICI, E. TOSCANO, F. VIOLA, *A Mesh-free Particle Method for Transient Full-Wave Simulation*, IEEE Transactions on Magnetics, vol. 43, pages 1333-1336, 2007.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Newton's problem of minimal resistance in the class of solids of revolution

Alena Aleksenko[1] **and Alexander Plakhov**[1,2]

[1] *Department of Mathematics, University of Aveiro, Portugal*

[2] *Institute of Mathematical and Physical Sciences, Aberystwyth University, UK*

emails: alena-aleksenko@rambler.ru, a.plakhov@ua.pt

Abstract

Newton's problem of the body of minimal aerodynamic resistance is traditionally stated in the class of *convex* axially symmetric bodies with fixed length and width. We will discuss the minimal resistance problem in the wider class of axially symmetric but *generally nonconvex* bodies.

Key words: Newton's problem, bodies of minimal resistance

1 Introduction

The Newton's problem of minimal resistance can be expressed as follows. A body is placed in a parallel flow of point particles. The density of the flow is constant, and velocities of all particles are identical. Consider the class of convex and axially symmetric bodies inscribed in a given right circular cylinder, where the symmetry axis of the body and the cylinder axis coincide and are parallel to the flow velocity. Each particle incident on the body makes an elastic reflection from its boundary and then moves freely again. The flow is very rare, so that the particles do not interact with each other. Each incident particle transmits some momentum to the body; thus, there is created a force of pressure on the body; it is called *aerodynamic resistance force*, or just *resistance*.

Newton described the body of minimal resistance in the class of admissible bodies specified above. The rigorous proof of the fact that the described body is indeed the minimizer was given two centuries later.

Since the early 1990s, there have been obtained new interesting results related to the problem of minimal resistance in various classes of admissible bodies [1]-[4]. In particular, there has been considered the wider class of convex (generally non-symmetric) bodies inscribed in a given cylinder. It was shown that the solution in this class exists and does not coincide with the Newton one. The problem is not completely solved till now. By removing both assumptions of symmetry and convexity, one gets the (even wider) class of bodies inscribed in a given cylinder. More precisely, a generic body from the class is a connected set with piecewise smooth boundary which is contained in the orthogonal cross section of the cylinder and satisfies a regularity condition to be specified below. In contrast to the class of convex and axis-symmetric bodies, the infimum of resistance here equals zero and we believe the infimum cannot be attained.

We suppose that the radius of the cylinder equals 1 and the height equals h, with h being a fixed positive number. The purpose of this work is to consider the forth possible case in [4] meaning the class of bodies of revolution but generally nonconvex. Let us consider plane $\mathbb{R}^2$ with coordinates x, z and lets consider the rectangle $[0,\,1] \times [0,\,h]$ and a set F, which is inscribed into that rectangle. In other words, F belongs to the rectangle and contains points of each of its side. Besides that the set F is closed, connected and has symmetry with order to the axe Oz, moreover it has piecewise smooth boundary and satisfy the property of the billiard scattering regularity in $\mathbb{R}^2 \setminus F$. This property means that for almost all $x \in [0,\,1]$ the movement of the billiard particle with coordinates $x(t) = x,\;\; z(t) = -t$ under $t \le -h$ is defined for all $t \in \mathbb{R}$. This particle makes a finite number of collisions with the boundary ∂F and then it moves freely with some velocity $v_F(x) \in S^1$. The mentioned regularity property includes also the condition that function $v_F(x) = (v_F^x(x),\, v_F^z(x))$ is a measurable function of x. Denote by $\mathcal{F}_h$ the set of all described sets F. Corresponding of the body Ω_F, which was obtained as a rotation of $F \in \mathcal{F}_h$ about the axis Oz, equals to $-2\pi\rho(0,0,R(F))$, where ρ is the flow density, and $R(F) = \int_0^1 (1 + v_F^z(x))\, x\, dx$. Our aim is to minimize resistance in the class of bodies of revolution $\Omega_F,\;\; F \in \mathcal{F}_h$; Or in other words to find $\inf_{F\in\mathcal{F}_h} R(F)$.

Denote $\mathcal{F}_h^{conv}$ a class of convex sets $F \in \mathcal{F}_h$. Upper part of the boundary of every set $F \in \mathcal{F}_h^{conv}$ could be considered as a graphic of some function $z = h - f_F(|x|)$, where $f_F(0) = 0,\;\; f_F(1) \le h$, function f_F is convex $[0,\,1]$ and monotone nondecrease. Let us define *modified scattering law* on the boundary of the set $F \in \mathcal{F}_h^{conv}$ in the following way. Correspondingly to this law, the particle initially moves as $x(t) = x \in [-1,\,1]$, $z(t) = -t$, but after collision it gets a velocity $\widehat{v}_F(x)$, parallel to the tangent to ∂F in the collision point: $\widehat{v}_F(x) = (\widehat{v}_F^x(x), \widehat{v}_F^z(x)) = (\operatorname{sgn} x, -f_F'(|x|))\,/\sqrt{1 + {f_F'}^2(|x|)}$. The corresponding functional has formulation $\widehat{R}(F) = \int_0^1 (1+\widehat{v}_F^z(x))\, x\, dx$. So we obtain that $\widehat{R}(F)$ determine the resistance of the convex symmetric body Ω_F in case of modified scattering law.

The following theorem allows to restrict the problem of minimization of resistance R in the class of bodies of revolution to the problem of minimization of $\widehat{R}$ in the class of *convex* bodies of revolution.

Theorem 1. $\inf_{F\in\mathcal{F}_h} R(F) = \inf_{F\in\mathcal{F}_h^{conv}} \widehat{R}(F)$.

This theorem immediately follows from the next two lemmas, proves of which we don't show here.

Lemma 1 *Let $F \in \mathcal{F}_h$. Note that convex hull* $\operatorname{conv} F$ *belongs to F_h^{conv}. It holds $\widehat{R}(\operatorname{conv} F) \le R(F)$.*

Lemma 2 *Let $F \in \mathcal{F}_h^{conv}$. There exists a sequence of nonconvex sets $F_n \in \mathcal{F}_h$ such that* $\lim_{n\to\infty} R(F_n) = \widehat{R}(F)$.

The search for the minimum of $\widehat{R}$ is restricted now to minimization of $\int_0^1 (1 - \frac{f'(x)}{\sqrt{1+f'^2(x)}})\, x\, dx$ in the class of convex and nondecreasing functions f defined on [0, 1], satisfying relations $f(0) = 0,\ f(1) \le h$. This minimization is done using Pontryagin's minimum principle; as a result we obtain the following theorem.

Theorem 2.

$$\inf_{F\in\mathcal{F}_h^{conv}} \widehat{R}(F) = \frac{1}{2} - \frac{1}{16}\left(8 - 2c^{2/3} - 3c^{4/3}\right)\sqrt{1-c^{2/3}} + \frac{3c^2}{16}\ln\left(\frac{1+\sqrt{1-c^{2/3}}}{c^{1/3}}\right), \quad (1)$$

where $c = c(h)$ is a unique solution of the equation:

$$h = \int_c^1 \sqrt{\left(\frac{x}{c}\right)^{2/3} - 1}\, dx = -\frac{3}{8}\left((c^{1/3} - 2c^{-1/3})\sqrt{1-c^{2/3}} - \ln\left(\frac{1+\sqrt{1-c^{2/3}}}{c^{1/3}}\right)\right). \quad (2)$$

Set $\widehat{F}_h \in \mathcal{F}_h^{conv}$, and is the minimizer of the functional $\widehat{R}$, is given by

$$\begin{cases} 0 \le z \le h, & \text{if } |x| \le c \\ 0 \le z \le h - \int_c^x \sqrt{\left(\frac{t}{c}\right)^{2/3} - 1}\, dt, & \text{if } c < |x| \le 1. \end{cases} \quad (3)$$

Denote $\mathcal{R}(h) := \inf_{F\in\mathcal{F}_h} R(F)$ and $\mathcal{R}_N(h) := \inf_{F\in\mathcal{F}_h^{conv}} R(F)$ are minimal values for our problem and for Newton one correspondingly. The following are valid expressions: $\mathcal{R}(0^+) = \frac{1}{2}\mathcal{R}_N(0^+) = 1/2;\ \mathcal{R}(h) = \frac{1}{4}\mathcal{R}_N(h)(1+o(1)) = \frac{27}{128}\frac{1+o(1)}{h^2}$ under $h \to +\infty$.

Acknowledgements

This work was supported by *Centre for Research on Optimization and Control* (CEOC) from the "*Fundação para a Ciência e a Tecnologia*" (FCT), cofinanced by the European Community Fund FEDER/POCTI, and by the FCT research project PTDC/MAT/72840/2006.

References

[1] G. Buttazzo, B. Kawohl, *On Newton's problem of minimal resistance*, Math. Intell. **15**, 7-12 (1993).

[2] F. Brock, V. Ferone, and B. Kawohl., *A symmetry problem in the calculus of variations.* Calc. Var. **4**, 593-599 (1996).

[3] T. Lachand-Robert and M. A. Peletier, *Newton's problem of the body of minimal resistance in the class of convex developable functions.* Math. Nachr. **226**, 153-176 (2001).

[4] A. Yu. Plakhov. *Newton's problem of the body of minimal resistance with a bounded number of collisions.* Russ. Math. Surv. **58** N°1, 191-192 (2003).

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Convergence and Stability of Iterative Refinement using Neville Elimination

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

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

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

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

Abstract

We provide sufficient conditions to ensure the convergence and stability of iterative refinement using Neville elimination for the resolution of linear systems of equations.

Key words: Iterative refinement, Neville elimination, Total positivity
MSC 2000: 65F05, 65F10

1 Introduction

Let us consider a linear system of equations $Ax = b$, with A a real nonsingular matrix. Solving this system with some direct method, in floating point arithmetic, we get an approximation $\widehat{x}^{(0)}$ to the solution. Iterative refinement is a well established and studied technique to improve the accuracy of the computed solution $\widehat{x}^{(0)}$ of the linear system $Ax = b$. For $k = 1, 2, \ldots$, the kth iteration of iterative refinement involves the following three steps:

1. Compute the residual $r^{(k)} = b - A\widehat{x}^{(k)}$ to get $\widehat{r}^{(k)}$

2. Solve the system $A\, y^{(k)} = r^{(k)}$ to get $\widehat{y}^{(k)}$

3. Update the solution $\widehat{x}^{(k+1)} = \widehat{x}^{(k)} + \widehat{y}^{(k)}$

For more details on this technique and its implementation see, for example, [7], [11] and [15].

The usual method to solve a linear system of equations $Ax = b$ is Gaussian elimination. So, in the literature it has been considered the study of iterative refinement using Gaussian elimination from several points of view (see [7], [11], [15] and [16]).

The main purpose of this work is to study the convergence and stability of iterative refinement using Neville elimination, which is an alternative procedure to Gaussian elimination to transform a square matrix A into an upper triangular matrix U. Neville elimination makes zeros in a column of the matrix A by adding to each row a multiple of the previous one. Here we only give a brief description of this procedure (for a detailed and formal introduction we refer to [9]). If $A \in \mathbb{R}^{n\times n}$, the Neville elimination procedure consists of at most $n-1$ steps:

$$A = A^{(1)} \to \widetilde{A}^{(1)} \to A^{(2)} \to \widetilde{A}^{(2)} \to \cdots \to A^{(n)} = \widetilde{A}^{(n)} = U.$$

On the one hand, $\widetilde{A}^{(t)}$ is obtained from the matrix $A^{(t)}$ by moving to the bottom the rows with a zero entry in column t, if necessary, to get that

$$\widetilde{a}_{it}^{(t)} = 0, \quad i \geq t \quad \Rightarrow \quad \widetilde{a}_{ht}^{(t)} = 0, \quad \forall\, h \geq i.$$

On the other hand, $A^{(t+1)}$ is obtained from $\widetilde{A}^{(t)}$ making zeros in the column t below the main diagonal by adding an adequate multiple of the ith row to the $(i+1)$th for $i = n-1, n-2, \ldots, t$. If A is nonsingular, the matrix $A^{(t)}$ has zeros below its main diagonal in the first $t-1$ columns. It has been proved that this process is very useful with totally positive matrices, sign-regular matrices and other related types of matrices (see [8] and [9]).

A real matrix is called totally positive (TP) if all its minors are nonnegative. TP matrices arise in a natural way in many areas of Mathematics, Statistics, Economics, etc. In particular, their application to approximation theory and Computer Aided Geometric Design (CAGD) is of great interest. For example, coefficient matrices of interpolation or least square problems with a lot of representations in CAGD (the Bernstein basis, the B-spline basis, etc.) are TP. Some recent applications of such kind of matrices to CAGD can be found in [12] and [14]. For applications of TP matrices to other fields see [8].

In [6], [9] and [10] it has been proved that Neville elimination is a very useful alternative to Gaussian elimination when working with TP matrices. In addition, there are some studies that prove the high performance computing of Neville elimination for any nonsingular matrix (see [3]). In [2] the backward error of Neville elimination has also been analyzed.

In [1] we give a sufficient condition that ensures the convergence of iterative refinement using Neville elimination for a system $A\,x = b$ with A any nonsingular matrix in $\mathbb{R}^{n\times n}$, and then we apply it to the case where A is TP. In this particular case, we refine here the sufficient condition for TP matrices of [1]. We prove that if

$$\gamma_{n-1}\|A\|\|A^{-1}\| < \frac{1}{2}, \tag{1}$$

where $\gamma_n := nu/(1-nu)$ and u is the unit roundoff, then iterative refinement using Neville elimination is convergent for any $\widehat{x}^{(0)}$.

We point out that the previous bound is of the same kind as those obtained by de Boor and Pinkus in [5] for Gaussian elimination.

Let us observe that the bound (1) depends on the condition number of the matrix A as every equivalent bound corresponding to Gaussian method. But, in contrast to most of the equivalent bounds for this method, bound (1) does not depend on the growth factor of the elimination procedure.

Another of the goals of this work is to study the stability of iterative refinement using Neville elimination. In Chapter 12 of [11] the stability of iterative refinement using Gaussian elimination was analyzed. Let us introduce now some basic concepts.

Given a linear system of equations $Ax = b$ with A a nonsingular matrix of order n, the componentwise backward error of an approximate solution $\hat{x}$ is defined as

$$\omega_{E,f}(\hat{x}) = \min\{\epsilon : \ (A+\delta A)\hat{x} = b + \delta b, \ |\delta A| \le \epsilon E, \ |\delta b| \le \epsilon f\} \tag{2}$$

where the matrix E and the vector f have nonnegative entries. We will adopt the usual choice for the tolerances E and f as

$$E = |A|, \quad f = |b|.$$

According to the result of Oettli and Prager (see Theorem 7.3 of [11] and [13]) the componentwise backward error can be computed in a simpler way than using formula (2) as

$$\omega_{|A|,|b|}(\hat{x}) = \max_i \frac{|r_i|}{(|A||\hat{x}| + |b|)_i}$$

where $r = b - A\hat{x}$ and with the subscript i we mean the ith component of the corresponding vector (the quotient $\alpha/0$ is interpreted as zero if $\alpha = 0$ and infinity otherwise).

An algorithm to solve a linear system of equations is said to be componentwise backward stable if the componentwise backward error is of order the unit roundoff provided that this is less than a certain threshold, that is to say

$$\omega_{|A|,|b|}(\hat{x}) = O(u) \quad \text{for} \quad u \le \bar{u}(n).$$

Nevertheless, for practical purposes, the threshold is also allowed to depend upon the data A and b, so that $\bar{u} = \bar{u}(n, A, b)$. Then, componentwise backward stability means that the approximate solution $\hat{x}$ is the exact solution of a perturbed system where the perturbations are small componentwise.

In the case of iterative refinement using Gaussian elimination Higham proved that under certain conditions the method is componentwise backward stable. Here we study the stability of iterative refinement when using Neville elimination instead of Gaussian elimination. So we have obtained a sufficient condition, which can be simplified for the particular case of TP matrices using the backward error analysis obtained in [2]. Specifically, for this sort of matrices we have seen that

$$|b - A\hat{x}^{(1)}| \leq 2\gamma_{n+1} A|\hat{x}^{(1)}|$$

provided that

$$\mathrm{cond}(A^{-1})\sigma(A, \hat{x}^{(1)}) \leq \frac{n+1}{2u\left(n+2+\frac{9\gamma_n}{4u}\right)^2},$$

where

$$\sigma(C, x) := \frac{\max_i(|C||x|)_i}{\min_i(|C||x|)_i}.$$

Then, by the result of Oettli and Prager, we can conclude that

$$\omega_{|A|,|b|}(\hat{x}^{(1)}) \leq 2\gamma_{n+1} = 2(n+1)u + O(u^2).$$

and, therefore, one step of iterative refinement using Neville elimination is backward stable.

Acknowledgements

This work has been partially supported by the Spanish Research Grant MTM2006-03388 and under MEC and FEDER Grant TIN2007-61273.

References

[1] P. Alonso, J. Delgado, R. Gallego and J. M. Peña, *Iterative Renement for Neville Elimination*, to appear in International Journal of Computer Mathematics.

[2] P. Alonso, M. Gasca and J. M. Peña, *Backward Error Analysis of Neville Elimination*, Appl. Numer. Math. **23** (1997) 193–204.

[3] P. Alonso, R. Cortina, I. Díaz and J. Ranilla, *Neville Elimination: a Study of the Efficiency Using Checkerboard Partitioning*, Linear Algebra Appl. **393** (2004) 3–14.

[4] T. Ando, *Totally Positive Matrices*, Linear Algebra Appl. **90** (1987) 165–219.

[5] C. de Boor and A. Pinkus, *Backward Error Analysis for Totally Positive Linear Systems*, Numer. Math. **23** (1997) 193–204.

[6] J. Demmel and P. Koev, *The Accurate and Efficient Solution of a Totally Positive Generalized Vandermonde Linear System*, SIAM J. Matrix Anal. Appl. **27** (2005) 142–152.

[7] G. E. Forsythe and C. B. Moler, *Computer Solutions of Linear Algebraic Systems*, Prentince-Hall, Englewood Cliffs, NJ, 1967.

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

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

[10] M. Gassó and J. R. Torregrosa, *A Totally Positive Factorization of Rectangular Matrices by the Neville elimination*, SIAM J. Matrix Anal. Appl. **25** (2004) 986–994.

[11] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[12] H. Lin, H. Bao and G. Wang, *Totally positive bases and progressive iteration approximation*, Comput. Math. Appl. **50** (2005) 575–586.

[13] W. Oettli and W. Prager, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math. **6** (1964) 405-409.

[14] J. M. Peña, *Shape preserving representations in Computer Aided-Geometric Design*, Nova Science Publishers, Inc., New York, 1999.

[15] J. H. Wilkinson, *Errors Analysis of Direct Methods of Matrix Inversion*, J. Assoc. Comput. Mach. **8** (1961) 281-330.

[16] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Notes on Applied Science **32** (Her Majesty's Stationery Office, 1963).

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

A decision-making differential model for social insects

Raul Abreu de Assis[1], **Ezio Venturino**[2] **and Wilson Castro Ferreira Jr**[3]

[1] *Departamento de Matemática, Universidade do Estado de Mato Grosso*

[2] *Dipartimento di Matematica, Università degli Studi di Torino*

[3] *Departamento de Matemática Aplicada, Universidade Estadual de Campinas*

emails: raulassis@yahoo.com, ezio.venturino@unito.it, wilson@ime.unicamp.br

Abstract

In this work we present a model for the phenomena of collective decision making in social insects using an n-dimensional system of differential equations. We perform a complete stability analysis for a special case in the model, and present numerical simulations to illustrate the behavior of the model in the more general case. The analysis shows that up to a range of values of the parameters in the model, different processes of decision-making in the social insects could be modelled by the same mathematical equations.

Key words: template, instructions
MSC 2000: AMS codes (optional)

1 Introduction

Modelling decision-making in colonies of social insects is an activity that can contribute at least in two different ways to the progress of science. Through the use of mathematical models, it can shed light on the biological mechanisms the species use to achieve an efficient behavior. On the other hand, the attempt of a mathematical reformulation of an efficient behavior, has led to the creation of new mathematical methods, for instance, ant optimization algorithms [5].

Also, the metaphor of the social insects, its robustness and parallel processing characteristics have inspired a different approach to the management of complex systems. This approach is often called “Swarm Intelligence” and another important feature in it is the absence of a central control [2]. The swarm intelligence approach has been used to design, among other applications, routing algorithms [4] and cooperating robots [7].

In this paper we are concerned with the “decision” made by a colony of social insects when confronted with several options. This could include collective nest choice [8, 9] by ants, construction through self-assembling [6] or choice of path to a food source [3].

The central idea is to develop a general model that pictures the competition of the different options of the colony for its limited resources. The two acting forces, as it is usual for mechanisms of self-organization, is a positive and a negative feedback. The positive feedback exists in the form that the more resources of the colony an option has attracted, the easier it gets for the option to attract even more resources. For instance, if an ant has to choose between two branches and one of them has higher concentration of pheromone, the chance that the ant will choose the higher concentration branch is greater than the chance of choosing the other. In this manner, a positive feedback is created. The negative feedback is due to the fact that resources decay at given rates. In the case of self-assembling agents trying to form a bridge, the individuals have a certain probability of giving up their task [6]. Then, as more agents are recruited for bridge building, the rate of individuals giving up the task also increases. Thus, when an option attracts more resources, the total amount of decaying resources increases.

At first we present a two-dimensional model to introduce the central aspects of the modelling and analysis. We then extend it to the n-option model of collective choice.

2 The bidimensional model

To model a general process of decision-making, we will think of the colony as having a type of resource that it can provide at a constant rate F. The type of resource depends on the specific biological case under study. For instance, if we are studying the choice of a new nest site for a colony, as in [8], the “resource” is represented by the scouts looking for new sites. On the other hand, when modelling a choice of path mediated by pheromone, the “resource” is the amount of pheromone provided by the scouts.

For this two-dimensional model we will also assume that the colony has two options to allocate its resources. These options may correspond to two different nest sites, branches, groups of workers trying to build a bridge, *etc.* We say that the colony has made a "decision" when it directs most of its resources to only one of its options. Hence the variables x and y represent, respectively, the quantity of resources allocated to the different options A and B available for the colony.

Finally, we add the hypothesis that the resources are lost, decaying from the options of the colony at a constant rate θ. For pheromone this would be analogous to evaporation, while in the self-assembling case it is analogous to the fraction of workers that give up the formation per time unit [6].

2.1 The feedback function

The next step is to model the fraction of the resources that is assigned to each of the options. We assume that the colony is able to provide the resources at a constant rate. As a self-assembled bridge of workers gets bigger, the chance of attracting new workers to it increases [6], a branch having a higher pheromone concentration compared to others has a higher possibility of attracting scouts [3] and a nest site having more scouts on it than other sites has an increased chance of attracting even more scouts [8, 9]. To simulate this kind of behavior, we can use a threshold-type function to model the fraction of resources assigned to option A of the form

$$P_A(x,y) = \frac{(x+k)^p}{(x+k)^p + (y+k)^p}. \tag{1}$$

This particular type of threshold-type function has been used with success to model both task allocation in ant colonies and the choice of path to a food source [2]. The biological interpretation of the parameters is as follows

- $p \in \mathbf{R}^+$ represents the inherent sensibility of the agents (social insects) in perceiving concentration differences of resources. To see it clearly one has just to imagine that if x is just a little bigger than y and p has a large value then $P_x \approx 1$ (one could take $x = y + \epsilon$ and take the limit as $p \to \infty$).

- $k \in \mathbf{R}^+$ is the parameter that regulates the *threshold* effect on the dynamics of the model. In the situation where the choice represents the choice of a path marked by pheromones, k is the quantity of pheromone needed to make the distribution of resources significantly asymmetric (for instance, if $x \ll k$ and $y \ll k$, $P_x \approx 1/2$). In the study of the choice of a new nest site, k might be interpreted to be correlated to the threshold *quorum* that the workers use [9] to start transportation to the new nest site. Low values for k imply that the threshold is reached sooner, leading to fast decision-making while high values lead to slow decision-making, and eventually, as the value increases, no decision at all (a situation where the colony distributes its resources among the options). In figure 1 we present graphics of the function $P_A(x,y)$ for different values of y and k.

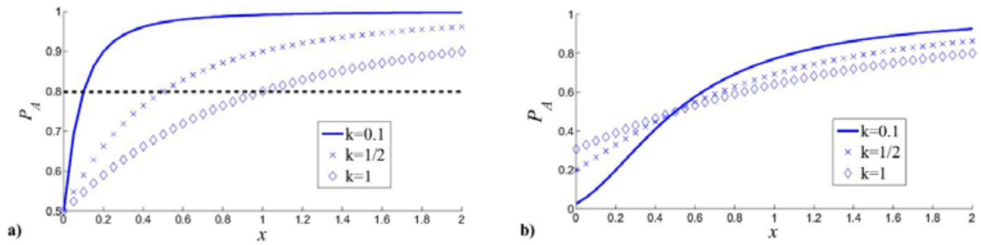


Figure 1: Graphics of the threshold function $P_A(x, y)$. **a)** Graph of P_A when $y = 0$. When $x = k$, the fraction of resources directed to option A reach 80%. **b)** Graph of P_A when $y = 0.5$.

2.2 The system

If x is the amount of colony resources allocated to the choice A and P_A is the fraction of the constant rate F of resources allocated to this choice, then the dynamics can be represented as

$$\frac{d\,x}{d\,t} = P_A F - \theta x. \tag{2}$$

This type of model would account for the dynamics of choosing between equal quality options. But in the process of decision-making, colonies often have to choose between options of different qualities. For instance, when choosing between different food sources, there is evidence (Sudd and Franks [10],p.114) that some ant species can modulate the amount of pheromone deposited in a trail depending on the quality of the food source. When confronted with several options for new nest sites, there is evidence [9] that the agents take more time to recruit nest mates for low-quality nest sites, leading to a lower rate of resources assigned to poorer choices.

These evidences lead us to the fact that the rate of distribution of the resources to a choice is dependent not only on the amount of resources assigned to it but also on the intrinsic quality of the choice. To model this behavior we use parameters to represent the quality of each different choice, α_A, α_B (we observe that units of α_A and α_B are the units of resources, while F (as θ) is measured in $[\text{time}]^{-1}$).

When this is incorporated into the model, we obtain the system of differential equations

$$\begin{aligned} \frac{dx}{dt} &= \frac{(x+k)^p}{(x+k)^p + (y+k)^p}\alpha_A F - \theta x \\ \frac{dy}{dt} &= \frac{(y+k)^p}{(x+k)^p + (y+k)^p}\alpha_B F - \theta y \end{aligned} \tag{3}$$

To make the analysis of this general model, we will start with a particular case, when the sensibility exponent p is equal to 2 and the threshold constant k is equal to zero. At first sight this may look as an artificial case, and even a mathematically inconvenient one, since it adds a singularity at the origin to the equations. But as we shall see, this particular case can give us all the insight necessary to grasp the behavior of the general model.

Applying the adimensionalization $x^* = x/\alpha_A$, $y^* = y/\alpha_A$ and $t^* = \theta t$ and dropping the stars

$$\begin{aligned} \frac{dx}{dt} &= \frac{x^2}{x^2+y^2} f - x \\ \frac{dy}{dt} &= \gamma \frac{y^2}{x^2+y^2} f - y \end{aligned} \tag{4}$$

where $f = F/\theta$ and $\gamma = \alpha_B/\alpha_A$ are adimensional parameters. The first parameter is a measure of the relationship between how fast the colony can provide resources and how fast they decay. The second parameter measures the ratio between the options quality ($\gamma > 1$ means that the option B is of superior quality, for instance).

2.3 Stability analysis

The stationary points of the system 4 are $P_0 = (x_0, y_0) = (f, 0)$, $P_1 = (x_1, y_1) = (0, \gamma f)$ and $P_2 = (x_2, y_2) = \left(\frac{\gamma}{\gamma^2+1} f, \frac{\gamma^2}{\gamma^2+1} f\right)$. The Jacobian matrix at a generic point (x, y) is given by

$$J(x,y) = \begin{bmatrix} \dfrac{2fxy^2}{(x^2+y^2)^2} - 1 & \dfrac{-2yx^2 f}{(x^2+y^2)^2} \\ \dfrac{-2xy^2\gamma f}{(x^2+y^2)^2} & \dfrac{2\gamma f y x^2}{(x^2+y^2)^2} - 1 \end{bmatrix} \tag{5}$$

It follows that the characteristic polynomials at the points P_0, P_1 and P_2 are respectively $p_0(\lambda) = p_1(\lambda) = (1+\lambda)^2$ and $p_2(\lambda) = (1+\lambda)(1-\lambda)$, leading to stability for P_0, P_1 and instability for P_2. It is easy to show [1] that there is a separatrix $y = x/\gamma$ of the region of attraction of P_0 and P_1 passing through P_2. Also if (x^*, y^*) is a stationary point for the system 4 then it is on the line $y = \gamma(f - x)$. These characteristics are shown in Figure 2.

2.4 Analyzing the parameters

When $k = 0$ the parameters F and θ do not affect the qualitative behavior of the mathematical model, as can be seen from the characteristic polynomials p_0, p_1 and p_2. The parameter γ has an important role, as it represents the relative qualities of the options of the colony. Changing the value of γ leads to different slopes for the separatix: $\gamma > 1$ means that option B has a better quality. In the model this is reflected by a greater region of attraction for P_1, meaning that the colony has a higher chance of choosing option B over option A. If $\gamma < 1$ the situation is reversed and when $\gamma = 1$ both regions of attractions have the same size, meaning that both options are equally attractive.

Now to understand the role of the parameter k in the model, we shall use the necessary condition that a stationary point lies on the line $y = \gamma(f - x)$. Using this condition we get $\dfrac{dy}{dx} = \dfrac{\dot{y}}{\dot{x}} = -\gamma$, meaning that a trajectory starting on the line remains on it.

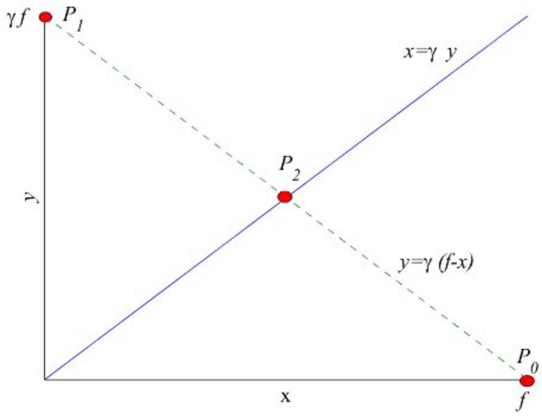


Figure 2: Stationary points of the bidimensional model. P_0, P_1 are stable and P_2 is unstable. The separatrix is defined by the equation $x = \gamma y$ and defines the region of attraction of P_1 and P_0. All stationary points are on the line $y = \gamma(f - x)$.

We can then write an equation for the dynamics of x

$$\frac{dx}{dt} = \frac{(x+k)^2\ f}{(\gamma\ (f-x)+k)^2 + (x+k)^2} - x = h(x)/s(x), \tag{6}$$

where

$$h(x) = -x(f-x)(\gamma^2 f - x\gamma^2 - x) - 2x(\gamma-1)(f-x)\ k + (f-2x)\ k^2, \tag{7}$$

$$s(x) = (\gamma(f-x)+k)^2 + (x+k)^2. \tag{8}$$

As $s(x) > 0,\ \forall x\ \in \mathbb{R}$, the stationary points on the line $y = \gamma(f-x)$ are defined by the roots of $h(x)$. When $k = 0$, we have the roots $x_0 = 0$, $x_1 = f$ and $x_2 = f\frac{\gamma^2}{\gamma^2+1}$ that are the stationary points for the system 4. Taking into account that k is a parameter, we can look at h as a continuous function of two variables $h \equiv h(x,k)$. The parameter k controls the location of the stationary points along the line $y = \gamma(f-x)$. As the value of k increases, two stationary points collapse and we get only one stable stationary point as in figure 3.

One interpretation that can be made about the role of the parameter k is that it can increase the efficiency of the colony in choosing the best option, since it gives more time to the colony for exploring the different options. In the case of nest-site choice, the parameter k may be interpreted as correlated with a *quorum*. In [8] when analyzing the choice between a superior and inferior nest site, labeled 2 and 1, respectively, the authors write:

> [...] At lower quorum sizes, transport begins to both sites, although it does sooner to site 2. A minority of the colony is thus carried to site 1 and

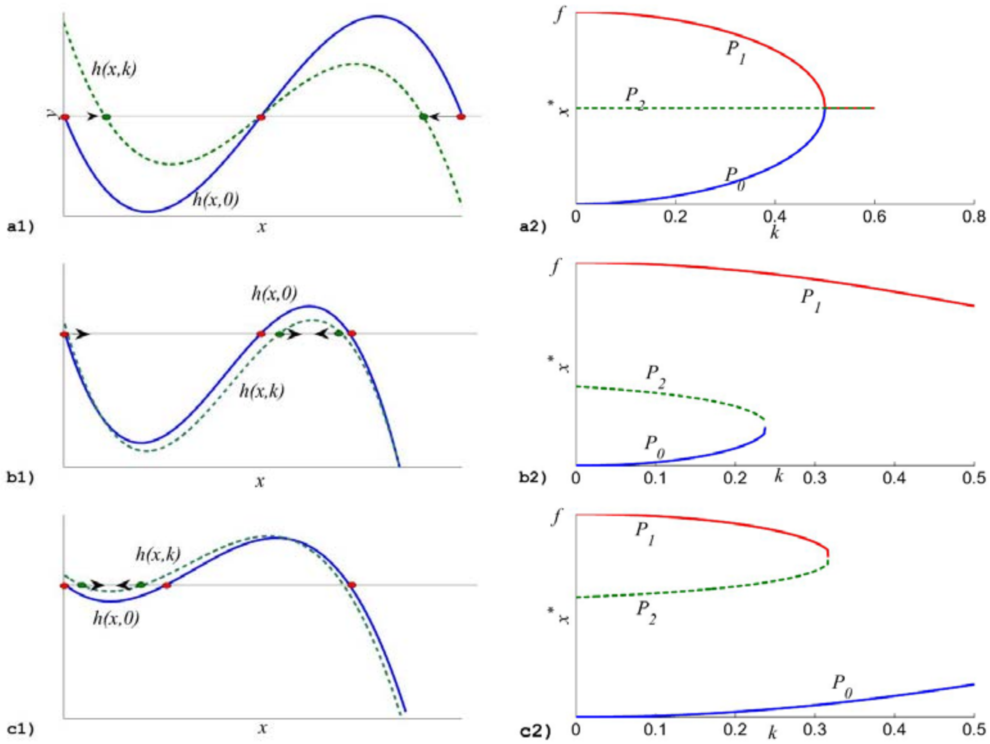


Figure 3: In the above graphics, $f = 1$, but different values lead to the same qualitative results. Left: graphics of the polynomials $h(x,k)$ on the line $y = \gamma(f-x)$ and the location of the stationary points as k varies. Right: trajectories of the x coordinates of the stationary points. **a)** $\gamma = 1$, x_1 and x_0 collapse into x_2. This represents a situation where the colony distributes its resources among the options. **b)** $\gamma < 1$, x_0 and x_2 collapse, being analogous to a situation where the colony chooses the option A. **c)** $\gamma > 1$, x_1 and x_2 collapse, being analogous to a situation where the colony chooses the option B.

> thus must be retrieved in a lengthy reunification phase. At higher quorum sizes, the better site maintains its advantage, but emigration time increases because colonies must spend longer in the slow tandem-run phase before initiating rapid transport.

Therefore, while increasing k leads to a higher chance of choosing the best option, it also delays the decision of the colony. In this model, k also affects the final distribution of resources between the options, being the distribution more even when k is high (when k is zero, all the resources are directed to only one option). It is important to notice that this qualitative discussion for k is related to the f value, but for fixed f we get the same qualitative behavior while varying the k parameter.

Another remark that must be made is that as f measures the abundance of the resources of the colony, for very low values of f (now k is fixed) there is no emergence of collective decision. This result can also can be identified in the behavior of colonies

[6]:

> [...]. A small population will never produce a chain, due to an incoming flow unable to compensate for the departure from the chains, characterized by a high leaving probability.[...]

The same qualitative behavior can be observed when a colony has to choose between two different paths to exploit a food source. Too few ants cannot produce an emergent choice of path as a minimum population size is needed to create a strong positive feedback.

The γ parameter has the clear role (as in figure 3) of determining which option will be chosen and also the fraction of resources attributed to the different options.

3 The n-dimensional model

The n-dimensional model represents the situation where the colony has to choose between n options. Now x_i is the amount of resources of the colony assigned to the option i, $\gamma_i = \alpha_1/\alpha_i$ are the parameters that measure the relative quality of the different options. Again we start the analysis looking at the case where $k = 0$, then the n-dimensional system of equations is

$$\frac{dx_i}{dt} = \frac{x_i^2}{\sum_{i=1}^{n} x_i^2}\gamma_i f - x_i \qquad\qquad i = 1, ..., n. \tag{9}$$

To make the analysis of the stationary points we first seek the points with all non-zero coordinates. Defining $L_n = \sum_{j=1}^{n} \frac{1}{\gamma_j^2}$, we have the stationary point

$$x_i^* = \frac{f}{\gamma_i L_n} \tag{10}$$

Let $P = (x_1^*, x_2^*, \ldots, x_n^*)$ and $\rho^2 = \sum_{j=1}^{n} x_j^2$. The n-dimensional Jacobian matrix is

$$J_n(\vec{x}) = \begin{bmatrix} \frac{2f\gamma_1 x_1(\rho^2 - x_1^2)}{\rho^4} - 1 & \frac{-2f\gamma_1 x_1^2 x_2}{\rho^4} & \dots & \frac{-2f\gamma_1 x_1^2 x_n}{\rho^4} \\ \frac{-2f\gamma_2 x_2^2 x_1}{\rho^4} & \frac{2f\gamma_2 x_2(\rho^2 - x_2^2)}{\rho^4} - 1 & \dots & \frac{-2f\gamma_2 x_2^2 x_n}{\rho^4} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-2f\gamma_n x_n^2 x_1}{\rho^4} & \frac{-2f\gamma_n x_n^2 x_2}{\rho^4} & \dots & \frac{2f\gamma_n x_n(\rho^2 - x_n^2)}{\rho^4} - 1 \end{bmatrix}. \tag{11}$$

The Jacobian at P is given by

$$J_n(P) = \begin{bmatrix} 1 - \frac{2}{L_n\gamma_1^2} & \frac{-2}{L_n\gamma_1\gamma_2} & \dots & \frac{-2}{L_n\gamma_1\gamma_n} \\ \frac{-2}{L_n\gamma_1\gamma_2} & 1 - \frac{2}{L_n\gamma_2^2} & \dots & \frac{-2}{L_n\gamma_2\gamma_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-2}{L_n\gamma_1\gamma_n} & \frac{-2}{L_n\gamma_2\gamma_n} & \dots & 1 - \frac{2}{L_n\gamma_n^2} \end{bmatrix} \tag{12}$$

and has the eigenvalue $\lambda_2 = 1$ with multiplicity $n-1$ associated with the eigenvectors $v_j = (0, ..., 1, ..., \gamma_n/\gamma_j)$, where the 1 is located in the j-th position, $j = 1, 2, ..., n-1$ and the eigenvalue $\lambda_1 = -1$ associated with the eigenvector $v_n = (1, \gamma_1/\gamma_2, \gamma_1/\gamma_3, \ldots, \gamma_1/\gamma_n)$. In view of the first positive eigenvalue, P is always unstable. Trajectories will be repelled away from it. They must then approach some other equilibrium, which can only be a boundary point in the n dimensional phase space.

To analyze the stability of these equilibria, observe that at least one of their components will vanish. Without loss of generality, let us take it to be x_n. It follows that $\dot{x}_n = 0$. Letting $\mathbf{0}_{n-1}$ denote the $n-1$ dimensional null vector, the Jacobian takes the form

$$A = \begin{bmatrix} J_{n-1} & \mathbf{0}_{n-1} \\ \mathbf{0}_{n-1}^T & -1 \end{bmatrix}. \tag{13}$$

The eigenvalues of the matrix A are -1 and the eigenvalues of J_{n-1}. Thus the stability analysis is exactly the same as for the stationary point P, with a reduction in the dimensionality (just the index changes to $n-1$). By repeating this argument inductively, the stability analysis for P will be used to deduce the stability of all the other points of the system.

The characteristic polynomial related to the equilibrium $P = P^{[n]}$ is

$$p^{[n]}(\lambda) = (\lambda - 1)^{n-1}(\lambda + 1).$$

Notice that the superscript is related to the point, while the degree is always n in λ. Allowing one variable to be zero, in view of (13) the characteristic polynomial becomes

$$p^{[n-1]}(\lambda) = (\lambda - 1)^{n-2}(\lambda + 1)^2,$$

which is again a polynomial of degree n in λ. Inductively, then

$$p^{[n-k]}(\lambda) = (\lambda - 1)^{n-k-1}(\lambda + 1)^{k+1}.$$

All these have at least one positive eigenvalue, except for $p^{[1]}(\lambda) = (\lambda + 1)^n$. Therefore the only stable stationary points are the ones in which only one variable is different from zero.

The parameters in the n-dimensional model have exactly the same interpretation as in the bidimensional one. We will show the bifurcation patterns for the tridimensional case, observing that the same patterns emerge for systems with higher dimensions.

For the tridimensional case, the stable stationary points are $P_1 = (\gamma_1 f, 0, 0)$, $P_3 = (0, \gamma_2 f, 0)$ and $P_5 = (0, 0, \gamma_3 f)$ and the unstable ones are:

$$\begin{aligned}
P_2 &= \left(\frac{f}{\gamma_1(1/\gamma_1^2+1/\gamma_2^2)}, \frac{f}{\gamma_2(1/\gamma_1^2+1/\gamma_2^2)}, 0\right) \\
P_4 &= \left(0, \frac{f}{\gamma_2(1/\gamma_2^2+1/\gamma_3^2)}, \frac{f}{\gamma_3(1/\gamma_2^2+1/\gamma_3^2)}, 0\right) \\
P_6 &= \left(\frac{f}{\gamma_1(1/\gamma_1^2+1/\gamma_3^2)}, 0, \frac{f}{\gamma_3(1/\gamma_1^2+1/\gamma_3^2)}\right) \\
P_7 &= \left(\frac{f}{\gamma_1(1/\gamma_1^2+1/\gamma_2^2+1/\gamma_3^2)}, \frac{f}{\gamma_2(1/\gamma_1^2+1/\gamma_2^2+1/\gamma_3^2)}, \frac{f}{\gamma_3(1/\gamma_1^2+1/\gamma_2^2+1/\gamma_3^2)}\right)
\end{aligned} \tag{14}$$

in figure 4 we show the bifurcation patterns when $f = 1$, the roots being approximated by numerical computations.

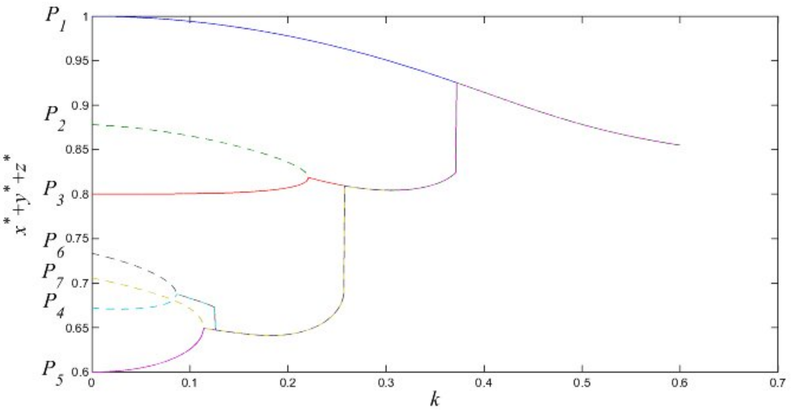


Figure 4: Evolution of the stationary points of the tridimensional model when $f = 1, \gamma_1 = 1, \gamma_2 = 0.8$ and $\gamma_3 = 0.6$. The y axis contains the sum of the coordinates of the stationary points. We observe that, as the value of the k increases the stable points relative to the poorer options (P_3 and P_5) collapse and only the better option is left (P_1).

4 Conclusions

The analysis of the model clarifies the role of the threshold factor in the decision-making process in colonies of social insects. Threshold values too low lead to quick decisions, but also to a higher chance of choosing poorer options while higher threshold values lead to a longer time for decision-making but to a higher possibility of choosing the best option. Beyond a certain value, the threshold factor can prevent the colony to build up a strong positive feedback, making it unable even to make a choice. This is the case when k is so high that even though the only stationary point that is left is the one of the best options, its coordinates do not show a clear choice of option, but rather a distribution of the resources of the colony among the options according to its quality.

The above considerations also hold for the parameter f, representing the flux of resources provided by the colony. A too low flux cannot make the necessary positive feedback for the emergence of a choice, while one that is too high can lead to premature options. Therefore, there is an interplay between the flux of resources from the colony and the threshold values. As the colonies of social insects are able to make efficient choices, we are led to believe that natural selection may be playing a role in “tuning” the threshold values for the species, making them able to make decisions in a trade-off between speed and efficiency.

With relation to optimization algorithms, the γ coefficients (along with f) play the role of making the algorithms “greedy”, while the k parameter is a measure of “freedom for search”. Greedy algorithms usually converge fast, but may converge to globally poor solutions. So the same interplay between fast convergence and global search that is observed in the social insects is repeated in the case of optimization algorithms.

Finally, we observe that the qualitative results of the model can be found in field and experimental observations, [7, 8]. Future work may include the study of the trade-off between fast decision-making and choosing the best option, by creating a fitness function that incorporates both factors and optimizing the parameters in the model. Also, it is possible to make the generalization for other values of the p exponent in the feedback function, although numerical simulations do indicate that the same qualitative results are to be observed.

5 Acknowledgments

The support of the WWS project of the University of Torino is thankfully acknowledged by the first author during his extended visit, at the time the research was carried out.

References

[1] R. A. de Assis and L. M. E. de Assis *Um modelo diferencial de recrutamento de formigas* Biomatemática **17** (2007), 35–46.

[2] E. Bonabeau, M. Dorigo and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, 1999.

[3] J. L. Denebourg , S. Aron , S. Goss e J. M. Pasteels, *The self-organizing exploratory pattern of the argentine ant*, J. of Insect Behavior, **3** (1989), 159–168.

[4] G. Di Caro and M. Dorigo *AntNet: Distributed Stigmergetic Control for Communications Networks* Journal of Artificial Intelligence Research **9** (1998), 317–365.

[5] M. Dorigo and G. Di Caro *Ant algorithms for discrete optimization* Artificial Life **5** (1999), 137–172.

[6] A. Lioni and J. L., Denebourg, *Collective decision through self assembling*, Naturwissenschaften **91** (2004) 237–241.

[7] A. Martinolli, A. J. Ijspeert and F. Mondana, *Understanding collective aggregation mechanisms: From probabilistic modelling to experiments with real robots* Robotics and Autonomous Systems, **29** (1999): 51–63.

[8] S. C. Pratt, E. B. Mallon, D. J. T. Sumper and N. R. Franks, *Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant Leptothorax albipennis*, Behav. Ecol. Sociobiol. **52** (2002) 117–127.

[9] S. C. Pratt, *Behavioral mechanisms of collective nest choice by the ant Temnothorax curvispinosus*, Insect. Soc. **52** (2005) 383–392.

[10] J. H. Sudd and N. R. Franks, *The Behavioural Ecology of Ants*, Chapman and Hall, Cambridge, 1987.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Action-Angle variables for a Manev system in a rotating reference frame

M. C. Balsas[1], **E. S. Jiménez**[1] **and J. A. Vera**[1]

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

emails: `mcarmen.balsas@upct.es`, `elena.jimenez@upct.es`, `juanantonio.vera@upct.es`

Abstract

In this paper we describe in some invariant manifolds of the motion the global phase portraits of a Manev system in a rotating reference frame. We obtain a complete topological classification of the different invariant sets of the phase flow for this problem. In particular, the Liouville-Arnold theorem has been used to do a particular analysis of the momentum map in its critical values. Finally, the action-angle variables and the regions where they can be defined are obtained. These variables allow us to calculate the modified Keplerian elements of this problem, useful to elaborate a perturbation theory. The results can be applied to study the Mercury precession using Classical Mechanics without using Relative Mechanics.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Initial guess of the solution of dynamic optimization of chemical processes

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

1 Introduction

Obtaining the accurate solution of optimal control problems is a crucial aspect in many areas of applied science. In this paper we shall focus especially on problems that arise in chemical engineering. There is a vast array of numerical methods and software packages for solving dynamic optimization or optimal control problems numerically, such as: SOCS, RIOTS_95, DIRCOL, MISER3, MINOPT, NDOT, DIDO, GPOCS, DYNOPT or Bryson's Matlab code. Unfortunately, these packages require an initial guess of the solution to start the iterations. Sometimes, in cases when convergence is not obtained, it is likely that the initial guess is such that convergence to the true solution is impossible. In such cases, users of these packages are accordingly advised to try different sets of initial guesses.

In this paper we shall concentrate on the special structure that appears in numerous problems of chemical reactors; more specifically, in the nonlinear Continuous Stirred Tank Reactor (CSTR). We shall present a very simple method to obtain an initial guess for the solution for this complex system. Moreover, we shall show that, for the chemical processes tested, the initial guess is very close to the solution and that our initial guess is attracted to a global minimum. We shall show that the theory allows us to address a wide range of problems: constrained, unconstrained, nondifferentiable, etc., while employing a very short computation time in all cases.

2 Mathematical Formulation

A Lagrange type Optimal Control Problem (OCP) can be formulated as follows:

$$\min_{\mathbf{u}(t)} I = \int_0^{t_f} F\left(t, \mathbf{x}(t), \mathbf{u}(t)\right) dt \tag{1}$$

subject to satisfying:

$$\dot{\mathbf{x}}(t) = f\left(t, \mathbf{x}(t), \mathbf{u}(t)\right) \tag{2}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{3}$$

$$\mathbf{u}(t) \in U(t),\ 0 \leq t \leq t_f \tag{4}$$

where I is the performance index, F is an objective function, $\mathbf{x} = (x_1(t), ..., x_n(t)) \in \mathbb{R}^n$ is the *state vector*, with initial conditions $\mathbf{x}_0$, $\mathbf{u} = (u_1(t), ..., u_m(t)) \in \mathbb{R}^m$ is the *control vector* bounded by $\mathbf{u}_{\min}$ and $\mathbf{u}_{\max}$, U denotes the set of admissible control values, and t is the operation time that starts from 0 and ends at t_f. The *state variables* (or simply the *states*) must satisfy the *state equation* (2) with given initial conditions (3). In this statement, we consider that the final instant is fixed and the final state is free. Let H be the Hamiltonian function associated with the problem

$$H(t, \mathbf{x}, \mathbf{u}, \lambda) = F\left(t, \mathbf{x}, \mathbf{u}\right) + \lambda \cdot f\left(t, \mathbf{x}, \mathbf{u}\right) \tag{5}$$

where $\lambda = (\lambda_1(t), ..., \lambda_n(t)) \in \mathbb{R}^n$ is called the *costate vector.* The classical approach involves the use of Pontryagin's Minimum Principle (PMP), which results in a two-point boundary value problem (TPBVP). In order for $\mathbf{u} \in U$ to be optimal, a nontrivial function λ must necessarily exist, such that for almost every $t \in [0, t_f]$

$$\dot{\mathbf{x}} = H_\lambda = f \tag{6}$$

$$\dot{\lambda} = -H_{\mathbf{x}} \tag{7}$$

$$H(t, \mathbf{x}, \mathbf{u}, \lambda) = \min_{\mathbf{v}(t) \in U} H(t, \mathbf{x}, \mathbf{v}, \lambda) \tag{8}$$

$$\mathbf{x}(0) = \mathbf{x}_0; \lambda(t_f) = \mathbf{0} \tag{9}$$

In this paper we deal with various chemical models whose dynamic equations present a particular structure (we present the two dimensional case for the sake of simplicity):

$$\min_{u_1(t)} I = \int_0^{t_f} F\left(x_1(t), x_2(t), u_1(t)\right) dt \tag{10}$$

$$\dot{x_1}(t) = f\left(x_1(t), x_2(t), u_1(t)\right) \tag{11}$$

$$\dot{x_2}(t) = f\left(x_1(t), x_2(t)\right) \tag{12}$$

The principal characteristic of this system is the absence of the control u_2 in equations (10-12). In several previous papers [1,2], the authors have presented a very simple method that is able to solve, for a known x_2, the problem formed by the equations (10-11). We now adapt this method to obtain an initial guess for the solution of the system (10-11-12).

The idea consists in constructing x_1 in an approximate and similar way to how it is constructed in [1,2] and in simultaneously constructing x_2 using Euler's (or Euler's improved) method in (12). In the discretization process, the values of x_2 obtained at the prior nodes are used to calculate x_1 at each node, and the values obtained for x_1 are used to calculate x_2. The method that we have developed to obtain x_1 is based on the use of an integral form of the Euler equation, combined with the simple shooting method. In the next section we shall see the excellent behavior of our approach by means of several examples.

3 Examples

We analyze three cases. In Example 3.1 we first consider the nonlinear CSTR as being unconstrained. In Example 3.2, we generalize the previous example, considering the constrained case with bounded control. Finally, in Example 3.3. we present a nondifferentiable case. We now present only the first case.

3.1 Unconstrained CSTR

Let us consider the system consisting of the dynamic optimization of a first-order irreversible chemical reaction carried out under non-isothermal conditions in a CSTR. The equations describing the chemical reactor are

$$\frac{dx_1}{dt} = -(2+u)(x_1+0.25) + (x_2+0.5)\exp(\frac{25x_1}{x_1+2}) \tag{13}$$

$$\frac{dx_2}{dt} = 0.5 - x_2 - (x_2+0.5)\exp(\frac{25x_1}{x_1+2}) \tag{14}$$

The control variable $u(t)$ represents the manipulation of the flow-rate of the cooling fluid. Here $x_1(t)$ represents the deviation from the dimensionless steady-state temperature, and $x_2(t)$ represents the deviation from the dimensionless steady-state concentration. In this section, we consider the case in which the control u is unbounded, and the initial conditions $x_1(0) = 0.09$ and $x_2(0) = 0.09$ are used. The optimal control problem is to determine u in the time interval $0 \leq t < t_f$ that will minimize the quadratic performance index

$$I = \int_0^{t_f} (x_1^2 + x_2^2 + 0.1u^2)dt \tag{15}$$

subject to the nonlinear dynamic constraints, where the dimensionless final time t_f is specified as 0.78.

Using a control vector iteration procedure, Luus and Cormack [3] showed that there exists a local optimum of $I = 0.244425$ and a global optimum of $I = 0.133094$. This optimal control problem provides a good test problem for optimization procedures and is a member of the list of benchmark problems [4]. It has been used by Luus [5] to evaluate his Iterative Dynamic Programming (IDP) algorithm, and by Luus and Galli [6] to examine the multiplicity of solutions. Ali et al. [7] solved this problem using eight stochastic global optimization algorithms, the results obtained varying between $I = 0.135$ and $I = 0.245$. The CPU time used was quite high, in some case more than 2382 s .

We apply our simple method and present the results below. The minimum value of $I = 0.1334$ was obtained very rapidly. The computation time for 15 iterations, with a discretization of 100 subintervals, was 2.5 s .

Our method presents numerous advantages: It is very easy to programme, the theory allows us to address a wide range of problems, the computation time is very short, the initial guess is very close to the solution, and the initial guess is attracted to a global minimum. The resulting initial guess for the optimal control policy is given in Figure 1 and for the state trajectories in Figure 2.

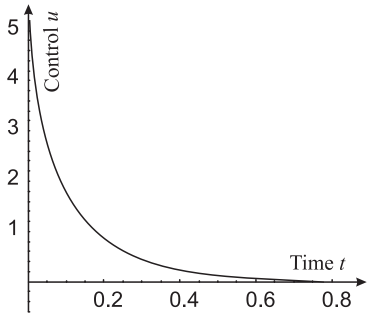


Fig. 1. Optimal control.

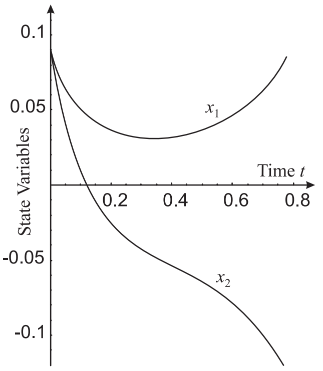


Fig. 2. Trajectories of the state variables.

References

[1] L. Bayon, J.M. Grau, M.M. Ruiz, and P.M. Suarez, *New Developments in the Application of Pontryagin's Principle for the Hydrothermal Optimization*, IMA Journal of Mathematical Control and Information **22(4)** (2005) 377-393.

[2] L. Bayon, J.M. Grau, M.M. Ruiz, and P.M. Suarez, *A Bolza Problem in Hydrothermal Optimization*, Applied Mathematics and Computation **184(1)** (2007) 12-22.

[3] R. Luus and D.E. Cormack, *Multiplicity of solutions resulting from the use of variational methods in optimal control problems*, Can. J. Chem. Eng. **50** (1972), 309-312.

[4] Ch.A. Floudas, et al., *Handbook of Test Problems in Local and Global Optimization*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[5] R. Luus, *Iterative Dynamic Programming*, Chapman & Hall/CRC Press, Boca Raton, FL, 2000.

[6] R. Luus and M. Galli, *Multiplicity of solutions in using dynamic programming for optimal control*, Hung. J. Ind. Chem. **19** (1991), 55-62.

[7] M.M. Ali, C. Storey, A. Torn, *Application of stochastic global optimization algorithms to practical problems*, J. Optim. Theory Applic. **95(3)** (1997) 545–563.

GLOBAL CONVERGENCE OF THE EPSILON STEEPEST DESCENT ALGORITHM

Rachid Benzine [1], Nacera Djeghaba [1]

[1] *Department of Mathematics,*
University Badji Mokhtar, B.P.12 Annaba, Algeria

emails: **Rachid.Benzine@univ-annaba.org**, **Nacera.Djeghaba@univ-annaba.org**

Abstract

Let : $f : R^n \rightarrow R$ and (P) the following problem of minimization without constraints:

$$(P): \quad \min\left\{f(x) : x \in R^n\right\}.$$

In [1], we introduced an algorithm so-called, the epsilon steepest descent, which accelerates the convergence of the gradient method. In [1] we have proved the global convergence of the epsilon steepest descent algorithm in the case of exact line searches.

In this work, we improve the algorithm introduced in [1] and we prove the global convergence of the new algorithm in the case of Armijo inexact line search.

Key words: Epsilon algorithm, steepest descent algorithm, global convergence

1. References

[1] **Djeghaba, N.**; **Benzine, R.** Acceleration de la convergence de la méthode de la plus forte pente. (French) [Acceleration of the convergence of the steepest descent method], *Demonstratio Math.* **39** (2006), no. 1, 169--181.

*Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008*

Cauchy boundary-conditioned Canonical formulations to physical problems simulations using arithmetical proprieties of the 4q-Boubaker polynomials

K. BOUBAKER[1] M. BOUHAFS[2] and O. Bamidele AWOJOYOGBE[3]

[1] ESSTT/Laboratoire de Physique de la Matière Condensée UPDS, Faculté des Sciences de Tunis, 2092 Tunis, TUNISIA.
[2] Unité de Recherche Mécanique Appliquée, Ingénierie et Industrialisation (MA2I).ENIT, TUNISIA.
[3] Department of Physics, Federal University of Technology, Minna, Niger-State, NIGERIA.

Abstract

In this paper, the arithmetical proprieties of the 4q-Boubaker polynomials subsequence are presented. In a context of applied physics, a polynomial expansion is presented with regard to particular boundary conditions. The latter form of this equation involves a well-known polynomial sequence: the Dickson polynomials. The arithmetical properties of these polynomials make the lastly proposed analytical solutions to some physics problems, like heat transfer in particular models, more appropriate to derive and to be involved in higher order systems.

M.S.C. 2000: 35K05, 33E30, 41A30, 41A55, 41A99.
PACS. 2008 : 02.00.00: Mathematical methods in physics - 02.30.Jr : Partial differential equations -02.30.Sa: Functional analysis - 47.10.A-: Mathematical formulations

1. Introduction

Recently, the approximation canonical schemes [1-5] which can yield satisfying solutions to the heat transfer, wave propagation and similar applied physics problems have been widely developed.

The use of polynomial expansions took a big part of these schemes and yielded meaningful results for both numerical and analytical analysis [6-9].

In this paper, we propose a canonical formulation to common physical problems with Cauchy boundary conditions. This formulation is based on demonstrated and verified arithmetical proprieties of a particular Boubaker polynomials subsequence of the: the 4q- Boubaker polynomials.

2. Historic of the Boubaker polynomials

2.a the Boubaker polynomials

The first monomial definition of the Boubaker polynomials[10-18] appeared in a physical study that yielded an analytical solution to heat equation inside a physical model[10]. This monomial definition is traduced by (eq. 1):

$$\underset{(1)}{B_n(X)} = \sum_{p=0}^{\xi(n)} \left[\frac{(n-4p)}{(n-p)} C_{n-p}^{p} \right] .(-1)^p .X^{n-2p}$$

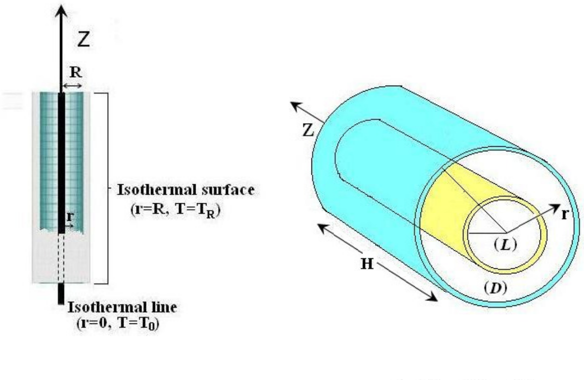
where:

$\xi(n) = \left\lfloor \frac{n}{2} \right\rfloor = \frac{2n+((-1)^n-1)}{4}$ (The symbol: $\lfloor \ \rfloor$ designates the Floor function)

The first few Boubaker polynomials are (eq.2):

$$\begin{aligned}
B_0(x) &= 1 \\
B_1(x) &= x \\
B_2(x) &= x^2 + 2 \\
B_3(x) &= x^3 + x \\
B_4(x) &= x^4 - 2 \\
B_5(x) &= x^5 - x^3 - 3x \\
B_6(x) &= x^6 - 2x^4 - 3x^2 + 2 \\
B_7(x) &= x^7 - 3x^5 - 2x^3 + 5x \\
B_8(x) &= x^8 - 4x^6 + 8x^2 - 2 \\
B_9(x) &= x^9 - 5x^7 + 3x^5 + 10x^3 - 7x \\
&\vdots
\end{aligned} \tag{2}$$

The Boubaker polynomials expansion method was used in the models (fig. 1) presented by works of O. Bamidele Awojoyogbe et *al.* in the field of organic tissues modelling[12], and the works of J. Ghanouchi et *al.* on the heat transfer modeling systems [11]. A part of these studies contains a conjoint work, based on the similarities between the hemodynamic flow system and a non proper differential equation attributed to the modified Boubaker polynomials.



a. Heat pipe model **b. Co-axial model**

Figure1: Studied models

A recent work presented by S. Slama et *al.*[19] presented a numerical model (fig. 2) of the spatial time-dependant evolution of A3 melting point in C40 steel material during a particular sequence of resistance spot welding.

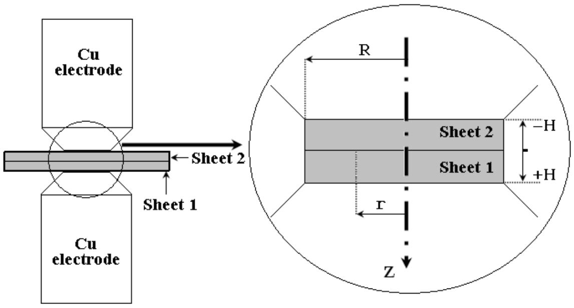
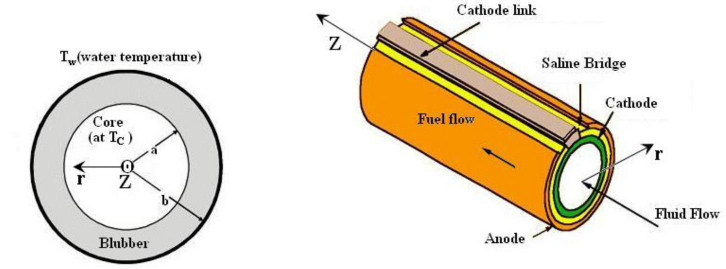


Figure 2: Spot welding model

The model was based on a solution to the heat equation, expressed, in cylindrical coordinates inside the joint area, in terms of pondered Boubaker polynomials expansions. Similar algorithm are being performed in the cases of some applied physics models (fig. 2).



a. Cetaciean body Transverse cut **b. Fuel-Air linear battery model**

Figure 3: Organic and industrial models

2.b the modified Boubaker polynomials (Boubaker-Turki polynomials)

The Boubaker-Turki polynomial or modified Boubaker polynomials[12-14], which are an enhanced form of the formerly defined polynomials, have been established as solutions to the second order differential equation (3):

$$(X^2-1)(3nX^2+n-2)\left[\tilde{B}_n(X)\right]'' + P_n(X)\left[\tilde{B}_n(X)\right]' + Q_n(X)\,\tilde{B}_n(X) = 0 \qquad (3)$$

where :

$$\begin{cases} P_n(X) = 3X(nX^2+3n-2) \\ Q_n(X) = -n(3X^2n^2+n^2-6n+8) \end{cases}$$

The modified Boubaker polynomials have a recursive coefficient definition expressed by equation (4):

$$\begin{cases} \tilde{B}_n(X) = \sum_{j=0}^{\xi(n)} \left[\tilde{b}_{n,j}\, X^{n-2j} \right]; \xi(n) = \dfrac{2n+((-1)^n-1)}{4} \\ \tilde{b}_{n,0} = 2^n; \qquad \tilde{b}_{n,1} = -(n-4)2^{n-2}; \\ \tilde{b}_{n,j+1} = \dfrac{(n-2j)(n-2j-1)}{(j+1)(n-j-1)} \times \dfrac{(n-4j-4)}{(n-4j)} \times \tilde{b}_{n,j} \\ \tilde{b}_{n,\xi(n)} = \begin{cases} (-1)^{\frac{n}{2}} \times 2 & \text{if n even} \\ 2(-1)^{\frac{n+1}{2}}(n-2) & \text{if n odd} \end{cases} \end{cases}$$

(4)

Both Boubaker and Boubaker-Turki polynomials are the source of several registered integer sequences [14,18].

The ordinary generating function of the Boubaker-Turki polynomials is (eq.5):

$$f_B(X,t) = \frac{1+3t^2}{1+t(t-2X)}$$

(5)

2.c The 4q-Boubaker polynomials subsequence

The Boubaker polynomials $B_n(X)$ explicit monomial form evoked, while prospected, some singularities for m=4, 8, 12, etc. In fact for the general case: m=4q the 2q rank monomial term is removed from the explicit form so that the whole expression contains only 2q effective terms. Correspondent 4q-order Boubaker polynomials[11] are presented in equation (6) as a general form and equation (7) as first functions:

$$B_{4q}(X) = 4\sum_{p=0}^{2q} \left[\frac{(q-p)}{(4q-p)} C_{4q-p}^{p} \right].(-1)^p.X^{2(2q-p)}$$

(6)

$$\begin{cases} B_0(\mathrm{X}) = 1; \\ B_4(\mathrm{X}) = X^4 - 2; \\ B_8(\mathrm{X}) = X^8 - 4X^6 + 8X^2 - 2; \\ B_{12}(\mathrm{X}) = X^{12} - 8X^{10} + 18X^8 - 35X^4 + 24X^2 - 2; \\ B_{16}(\mathrm{X}) = X^{16} - 12X^{14} + 52X^{12} - 88X^{10} + 168X^6 - 168X^4 - 48X^2 - 2; \\ B_{20}(\mathrm{X}) = X^{20} - 16X^{18} + 102X^{16} - 320X^{14} + 455X^{12} - 858X^8 + 1056X^6 - 495X^4 + 80X^2 - 2; \\ \dots\dots\dots\dots\dots\dots \end{cases} \quad (7)$$

3. The 4q-Boubaker polynomials subsequence proprieties

3.a Recursive relation

The main purpose of this section is to establish a recursive relation involving three consecutive 4q-Boubaker polynomials:

A conjectured primal form was (eq.8) :

$$B_{4(q+1)} = P(X) \times B_{4(q)} + Q(X) \times B_{4(q-1)} \quad (8)$$

Using the first terms of the 4q-Boubaker polynomials we obtain the empirical solution (9):

$$\begin{cases} P(X) = X^4 - 4X^2 + 2 = B_4^*(X) \\ Q(X) = -1 \end{cases} \quad (9)$$

We conjecture hence that we have the recursive relation (10):

$$B_{4(q+1)} = (X^4 - 4X^2 + 2) \times B_{4(q)} - B_{4(q-1)} = B_4^*(X) \times B_{4(q)} - B_{4(q-1)} \quad (10)$$

3.b Quadratic relation

Let's define the q-dependent function H_q :

$$H_q = B_{4q}^2(X) - B_{4(q-1)}(X) \times B_{4(q+1)}(X) \quad (11)$$

By calculating the difference (12):

$$H_{q+1} - H_q = B^2_{4(q+1)}(X) - B_{4q}(X)B_{4(q+2)}(X) - B^2_{4q}(X) + B_{4(q-1)}(X)B_{4(q+1)}(X) = 0$$
(12)

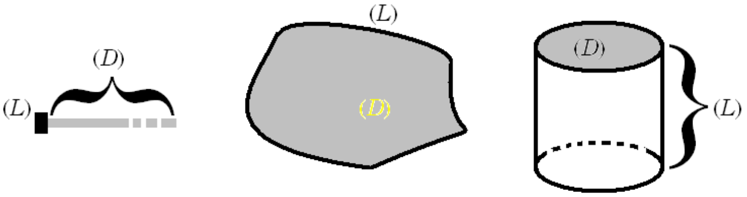
We demonstrate that H_j(X)is not q-dependent. Using the first terms of the 4q-Boubaker polynomials we obtain finally the quadratic relation (13):

$$B^2_{4q}(X) - B_{4(q-1)}(X) \times B_{4(q+1)}(X) = X^2(X^2-1)^2(3X^2+4) = B^*_8(X); \quad \forall q > 1 \qquad (13)$$

4. Cauchy boundary-conditioned Canonical formulations

4.a Cauchy boundary-conditions

When a differential equation, like a d'Alembert one, has to be solved inside a domain (*D*) contained inside a continuous closed boundary (*L*); the Cauchy boundary conditions[20-21] may be imposed. These conditions introduce two equations to be satisfied in boundaried (fig. 4. a and 4.b) by both the solution and its first derivative.



a. Linear model

b. 2-D model

c. Cylindrical model

Figure 4: Studied domains

An example is given in the case of periodic elliptic-parabolic type differential equation (14).

$$\begin{cases}\dfrac{\partial f(x,t)}{\partial t}-\nabla^2 f(x,t)=S(x,t);x\in(D)\\ f(x,t)=g(x,t);x\in(L),t>0\\ \dfrac{\partial f(x,t)}{\partial x}=h(x,t);x\in(L),t>0\end{cases}\tag{14}$$

where g and h are given functions, and S is the source term.

4.b Canonical 4q-Boubaker polynomials expansion

We consider, for example that (D) is a cylindrical volume(fig. 4.c). We have hence, in cylindrical coordinates (r,θ,z):

$$\begin{cases}(D)=\begin{cases}0<r<R\\0<z<H\\\theta\in[0,2\pi]\end{cases};(L)=\begin{cases}r=R\\0<z<H\\\theta\in[0,2\pi]\end{cases}\end{cases}\tag{15}$$

The 4q-Boubaker polynomials expansion is used in the case of a pulsed elliptic-parabolic problem with cylindrical symmetry. The solution of such problem is a spatial r-dependent distribution associated to a t-dependent source term modulated at a pulsation ω. The problem is traduced by the system (15):

$$\begin{cases}j\omega\times f(r)-\nabla^2 f(r)=S(r);r\in(D)\\ f(r)=g(r);r\in(L)\\ \dfrac{\partial f(r)}{\partial r}=h(r);r\in(L)\end{cases}\tag{16}$$

We consider, for example that (D) is a cylindrical unlimited volume. We have hence for an r-symmetrical problem :

$$\begin{cases}(D)=\begin{cases}0<r<R\\-\infty<z<+\infty\end{cases};(L)=\begin{cases}r=R\\-\infty<z<+\infty\end{cases}\end{cases}\tag{17}$$

We have also the boundary conditions (18):

$$\begin{cases}f(r)=0;r\in(L) & f(r)\neq 0;r=0.\\ \dfrac{\partial f(r)}{\partial r}=k=cst.;r\in(L) & \dfrac{\partial f(r)}{\partial r}=0;r=0.\end{cases}\tag{18}$$

The canonical 4q-Boubaker polynomials expansion is performed by the expression (19):

$$f(r)=T_0\times\lim_{N\to+\infty}\left[\frac{1}{2N}\sum_{q=1}^{N}\xi_q.B_{4q}(r\frac{\alpha_q}{R})\right]\tag{19}$$

Where α_q are the minimal positive roots of the Boubaker 4q-order polynomials B_{4q}, and ξ_q are coefficients to be found.

4.c Cauchy-related features

Due to the proprieties of the 4q-Boubaker polynomials (20), boundary conditions (18) are intrinsically satisfied.

$$\begin{cases} \left.\sum_{q=1}^{N}\xi_q .B_{4q}(r)\right|_{r=0} = 2N \neq 0; & \left.\sum_{q=1}^{N}\xi_q .\frac{\partial B_{4q}(r)}{\partial r}\right|_{r=0} = 0 \\ \left.\sum_{q=1}^{N}\xi_q .B_{4q}(\alpha_q (r/R))\right|_{r=R} = \left.\sum_{q=1}^{N}\xi_q .B_{4q}(\alpha_q)\right|_{r=R} = 0; & \left.\sum_{q=1}^{N}\xi_q .\frac{\partial B_{4q}(\alpha_q (r/R))}{\partial r}\right|_{r=R} \neq 0 \end{cases}$$

(20)

The main system (16); taking in account the Cauchy boundary conditions, is reduced to (21):

$$\begin{cases} \left. j\omega\times\frac{1}{2N}\sum_{q=1}^{N}\xi_q .B_{4q}(r\frac{\alpha_q}{R}) - \left(\frac{\alpha_q}{R}\right)^2\sum_{q=1}^{N}\xi_q .\frac{\partial^2 B_{4q}(r\frac{\alpha_q}{R})}{\partial r^2} = S(r)\right|_{N\to\infty} \quad ; r\in(D) \\ \sum_{q=1}^{N}\xi_q = 1 \end{cases} \tag{21}$$

The solution of the system (21) is achieved once the source term is decomposed as a similar expansion. The relevant task is hence to choose the source term the way it can be expressed by (22); or to expand the given source term in an equivalent form:

$$S(r) = \lim_{N\to+\infty}\left[\frac{1}{2N}\sum_{q=1}^{N} b_q .B_{4q}(r\frac{\alpha_q}{R})\right] \tag{22}$$

with b_q determined coefficients.

5. Conclusion

The proposed canonical expansion presents a supply to works aiming to solve applied physics equations. In fact, by the mean of the arithmetical proprieties of the 4q-Boubaker polynomials, the Cauchy boundary conditions are substantially verified and indirectly taken in account inside the main equation.

The canonical expansion presented in this paper, is being tested inside several actual applied physic models.

References

[1] A. Barinka, T. Barsch, P. Charton, A. Cohen, S. Dahlke, W. Dahmen, and K. Urban, Adaptive wavelet schemes for elliptic problems: Implementation and numericalexperiments, SIAM J. Sci. Comput. **23** (2001), no. 3, 910–939.

[2] S. Dahlke, M. Fornasier, and T. Raasch, Adaptive frame methods for elliptic operator equations, Adv. Comput. Math. **27** (2007), no. 1, 27–63.

[3] S. Dahlke, M. Fornasier, T. Raasch, R. Stevenson, and M. Werner, Adaptive frame
methods for elliptic operator equations: The steepest descent approach, IMA J. Numer. Anal. 27 (2007), no. 4, 717–740.

[4] R. P. Agarwal and G. V. Milovanovic, Extremal problems, inequalities, and classical orthogonal polynomials, Applied Mathematics and Computation, Volume **128**, Issues 2-3, 25 May 2002, Pages 151-166.

[5] N. Weyrich, Spline wavelets on an interval, Wavelets and Allied Topics, 2001, pp. 117–189.

[6] L. Rebillard and A. Ronveaux, Expansion of multivariate polynomials in products of univariate orthogonal polynomials: discrete case,Journal of Computational and Applied Mathematics, Volume **133**, Issues 1-2, 1 August 2001, Pages 567-578.

[7] C. Jungemann and B. Meinerzhagen, A Legendre Polynomial Solver for the Langevin Boltzmann Equation **3**, Numbers 3-4, 157-160 (2004).

[8] J. Waldvogel, Fast construction of the Fejér and Clenshaw-Curtis quadrature rules, BIT Numerical Mathematics 43 (1), p. 001-018 (2004).

[9] H.M. Srivastava**,** Certain q-Polynomial Expansions for Functions of Several Variables , IMA Journal of Applied Mathematics 1983 **30**(3),Pages 315-323.

[10] A. Chaouachi, K. Boubaker, M. Amlouk and H. Bouzouita, *Enhancement of pyrolysis spray disposal performance using thermal time-response to precursor uniform deposition*, Eur. Phys. J. Appl. Phys. **37** (2007) pp.105-109.

[11] J. Ghanouchi, H. Labiadh and K. Boubaker, An Attept to solve the heat transfer equation in a model of pyrolysis spray using 4q-Order Boubaker polynomials, Int. J.of Heat & Tech., in press.

[12] O. B. Awojoyogbe and K. Boubaker, A solution to Bloch NMR flow equations for the analysis of homodynamic functions of blood flow system using m-Boubaker polynomials, International Journal of Current Applied Physics, Elsevier, DOI : 10.1016/j.cap.2008.01.0193.

[13] K. Boubaker, *On modified Boubaker polynomials: some differential and analytical properties of the new polynomials issued from an attempt for solving bi-varied heat equation***.** Trends in Applied Science Research, by Academic Journals, '*aj*' New York; USA, ISSN : 1819-3579. 2(6) :540-544,2007.

[14] The Boubaker-Tuki polynomials (or Modified Boubaker polynomials), *Planet-Math Encyclopedia, The Mathematics worldwide Encyclopedia* (available also online at :
http://planetmath.org/encyclopedia/BoubakerTurkiPolynomials.html

[15] H. Labiadh et al. : *A Sturm-Liouville shaped characteristic differential equation as a guide to establish a quasi-polynomial expression to the Boubaker polynomials*, Journal of Differential equations and control processes, № 2, (2007),reg. № P2375, ISSN 1817-2172.

[16] O.T.P.D.A., *Les Polynômes de Boubaker*. Dépôt légal N°: 21-01-04-04-2007.Tunis, Tunisia

[17] K. Boubaker**,** *Les Polynômes de Boubaker, Une classe polynomiale qui émane d'un essai de résolution de l'équation de la chaleur,* Deuxièmes Journées Méditerranéennes de Mathématiques Appliquées JMMA02, Monastir,TUNISIE. March (2007).

[18] The Boubaker polynomials, *Planet-Math Encyclopedia, The Mathematics worldwide Encyclopedia* (available also online at : **http://planetmath.org/encyclopedia/BoubakerPolynomials.html**

[19] S. Slama, J. Bessrour, K. boubaker and M.Bouhafs, Investigation of A3 point maximal front spatial evolution during resistance spot welding using 4q-Boubaker polynomial sequence, *Proceedings of COTUME 2008*, pp 79 :80.(2008)

[20] H. Takaoka and Y. Tsutsumi, Well-posedness of the Cauchy problem for the modified KdV equation with periodic boundary condition, International Mathematics Research Notices. **56**, (2004), pp. 3009-3040.

[21] Y. L. Liang and H. Q. Chun **,** The Cauchy Boundary Value Problems on Closed Piecewise Smooth Manifolds in C^n, **20**, Number 6 (2004),pp : 989-998.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008

A Quantum Mechanical Description of the Laws of Relativity

Erkki J. Brändas

Department of Quantum Chemistry, Uppsala University, Uppsala, Sweden

email: Erkki.Brandas@kvac.uu.se

Abstract

We have examined the old purported dilemma of quantum mechanics versus the theory of relativity. By proposing a first principles, relativistically invariant theory, via an analytic extension of quantum mechanics into the complex plane we offer a model that (i) include features such as time- and length-scale contractions and (ii) suggest incorporation of gravitational interactions, (iii) the Einstein general relativistic law of light deflection and (iv) the compatibility with the Schwarzschild metric in a spherically symmetric static vacuum. The present viewpoint asks for a new perspective on the age-old problem of quantum mechanics versus the theory of relativity as the relation with the Klein-Gordon-Dirac relativistic theory confirms some dynamical features of both the special and the general relativity theory.

Keywords: Klein-Gordon-, Dirac equation, particle-antiparticles, complex symmetry, non-positive metric, Jordan blocks, special- and general relativity.
MSC2000:

1. Introduction

It is well-known that that the Klein-Gordon-Dirac equation can be written formally as a standard self-adjoint secular problem based on the simple Hamiltonian matrix (in mass units)

$$\boldsymbol{H} = \begin{pmatrix} m_0 & p/c \\ p/c & -m_0 \end{pmatrix}. \tag{1}$$

Here m_0 is the rest-mass, $p = mv$ is the momentum of the particle and c the velocity of light – note also that the entities in Eq. (1) are operators and the velocity v of the particle(s) is relative a system in rest, wherever the rest masses of the particles involved are m_0 and $-m_0$ respectively. Similar treatments can also be made in relation to the Dirac equation [1].

As shown elsewhere [2] an analogous formulation follows from the modification of considering a complex symmetric ansatz. This permits an important generalization since it will simultaneously allow the introduction of time- and length scales as well as mimic the non-positive definiteness of the Minkowski metric [3].

In this abstract we will briefly review the model. We will further discuss the extension to the general case of gravitational interactions. As a result we will connect with some of the most well known facts of the laws of special and general relativity, i.e. the contraction of length- and time scales, Einsteins law of light deflection in a gravitational field and the appearance of the Schwarzschild radius in a static symmetric vacuum. Finally the compatibility of the formulation with the Schwarzschild gauge in the minimal two-component metric is indicated.

The conclusions made suggest the proposition that “the Einstein laws of relativity” indeed is a quantum effect [1].

2. The complex symmetric ansatz

As presented elsewhere [1-3] we will set up a simple 2×2 complex symmetric matrix that (without interaction) displays perfect symmetry between the states of the particle and its antiparticle image[2].

$$\boldsymbol{H} = \begin{pmatrix} m & -iv \\ -iv & -m \end{pmatrix}. \qquad (2)$$

In Eq.(2) the diagonal elements are the energies associated with a particle with mass m in a state with wave vector $|m\rangle$ and the antiparticle state, assigned a negative energy $-m$ with the state vector $|\overline{m}\rangle$. For the case of fermions we will need the Dirac equation, see Ref.[1] for a detailed treatment. $-iv$ is the complex symmetric interaction, see below; and the minus sign is by convention. For zero interaction the diagonal elements are $\pm m_0$. Note that he vectors $|m_0\rangle$ and $|\overline{m}_0\rangle$can be chosen orthonormal, while $|m\rangle$ and $|\overline{m}\rangle$ in general are bi-orthogonal.

Solving the secular equation corresponding to the ansatz (2) one obtains the roots $\lambda_{\pm} = \pm m_0$ from $\lambda^2 = m_0^2 = m^2 - \nu^2 = m^2 - p^2c^{-2}$, admitting the kinematic perturbation $\nu = p/c$. Hence $m^2c^4 = m_0^2c^4 + p^2c^2$. In passing one should note that $p = mv$, with appropriate modifications for a particle in an electromagnetic or other field [3], is an operator, which in its extended form may not be self-adjoint. Besides getting back a Klein-Gordon type equation, one also attains the "eigensolutions"

$$\begin{aligned} |m_0\rangle &= c_1|m\rangle + c_2|\overline{m}\rangle;\ \lambda_+ = m_0; \quad |m\rangle = c_1|m_0\rangle - c_2|\overline{m}_0\rangle; \\ |\overline{m}_0\rangle &= -c_2|m\rangle + c_1|\overline{m}\rangle;\ \lambda_- = -m_0; \quad |\overline{m}\rangle = c_2|m_0\rangle + c_1|\overline{m}_0\rangle; \end{aligned} \tag{3}$$

with

$$c_1 = \sqrt{\frac{1+X}{2X}};\ c_2 = -i\sqrt{\frac{1-X}{2X}};\ m = \frac{m_0}{X};\ c_1^2 + c_2^2 = 1\ \ X = \sqrt{1-\beta^2};\ \ \beta = p/mc. \tag{4}$$

For "classical particles" we recover the familiar β factor, e.g. $p/mc = \upsilon/c$. In general we need to keep the order of the operators appearing in the secular equation although the present formulation is somewhat unspecified for simplicity. Since we respect complex symmetry our model admits, under suitable environmental interactions and/or correlations primary complex resonance energies commensurate with rigorous mathematics and precise boundary conditions [4]. Hence we find that

$$m_0c^2 \rightarrow m_0c^2 - i\frac{\Gamma_0}{2};\quad \tau_0 = \frac{\hbar}{\Gamma_0};\ mc^2 \rightarrow mc^2 - i\frac{\Gamma}{2};\ \tau = \frac{\hbar}{\Gamma}, \tag{5}$$

where Γ, τ and Γ_0, τ_0 are the half widths and lifetimes of the state respectively, and $\hbar$ is Planck's constant divided by 2π. Inserting (5) into our model above and separating real and imaginary parts one gets immediately the contractions

$$\Gamma_0 = \Gamma\sqrt{1-\beta^2}\,;\ \tau = \tau_0\sqrt{1-\beta^2}\ . \tag{6}$$

Comparing times in the two scales enforcing Lorentz-invariance for the length *l*, one finds directly

$$l = \frac{l_0}{\sqrt{1-\beta^2}};\ t = \frac{t_0}{\sqrt{1-\beta^2}};\ m = \frac{m_0}{\sqrt{1-\beta^2}}. \tag{7}$$

From the present development we see that the laws of special relativity appears as a consequence of the quantum mechanical superposition principle. In the next section we will extend the discussion by considering gravitational interactions.

3. The general case

We extend the model to include gravity by augmenting the present development in the basis $|m,\overline{m}\rangle$:

$$\boldsymbol{H} = \begin{pmatrix} m(1-\kappa(r)) & -i\nu \\ -i\nu & -m(1-\kappa(r)) \end{pmatrix}$$

$$\lambda^2 = m^2(1-\kappa(r))^2 - p^2/c^2;\ \ \lambda = m_0(1-\kappa(r));\ \ \nu = p/c \tag{8}$$

with

$$\kappa(r) = \mu/r;\ \ \mu = \frac{G \cdot M}{c^2}. \tag{9}$$

Here μ is the gravitational radius, G the gravitational constant, M a "classical mass" (which does not change sign when $m \to -m$) and $\nu = p/c$ as before. Also $\kappa(r) \geq 0$ depends on the coordinate r of the particle m, with origin at the center of mass of M. The coordinate r (and t) refers to a flat Euclidean space and the emerging scales define the curved space-time. Eq. (8) has the eigenvalues $\lambda_\pm$

$$\begin{gathered} m_0^2 = m^2 - p^2/(1-\kappa(r))^2 c^2;\ \lambda_\pm/(1-\kappa(r)) = \pm m_0 = \pm\sqrt{m^2 - p^2/(1-\kappa(r))^2 c^2} \\ m = m_0/\sqrt{1-\beta'^2};\ \beta' \leq 1;\ 1 > \kappa(r);\ \beta' = p/mc(1-\kappa(r)) = \upsilon/c(1-\kappa(r)). \end{gathered} \tag{10}$$

It can be shown [2] that the angular momentum, $m\upsilon r$, for a particle under the influence of a central force is a constant of motion and hence we obtain (here $m = \hat{m}_{op}$ has the eigenvalue m_0) the relation

$$m_0 \upsilon r = m_0 c\mu;\ \upsilon = \kappa(r)c = \mu c/r. \tag{11}$$

where the constant have been evaluated at the limiting velocity c and the limiting distance, the gravitational radius. It follows for a particle with a non-zero mass that a degeneracy (Jordan block) will occur for the Schwarzschild radius at $r=R_{LS}$, provided the mass M is entirely localized inside the sphere, i.e.

$$\tfrac{1}{2}m = m\upsilon/c = m\kappa(r);\ r = R_{LS} = 2\mu \tag{12}$$

It follows further that either we find $m \to \infty$ with m_0 finite or m is finite with $m_0 \to 0$. For instance in the former case at $r=R_{LS}$ Eq. (8) yields

$$\begin{gathered} \boldsymbol{H}_{deg} = \frac{1}{2}\begin{pmatrix} m & -im \\ -im & -m \end{pmatrix} \to \boldsymbol{H}_{deg} = \begin{pmatrix} 0 & m \\ 0 & 0 \end{pmatrix} \\ |0\rangle = \frac{1}{\sqrt{2}}|m\rangle - i\frac{1}{\sqrt{2}}|\overline{m}\rangle \\ |\overline{0}\rangle = \frac{1}{\sqrt{2}}|m\rangle + i\frac{1}{\sqrt{2}}|\overline{m}\rangle \end{gathered} \tag{13}$$

explicitly displaying a Jordan block (see e.g. [5,6] for definitions) structure at the singularity. Thus in summary the present model entails that a quantum particle will occupy one of two possible states. The identification of these states occurs through the interaction $\boldsymbol{v}$ and the emergence of the length and time scale contractions. For zero rest-mass particles we find, see [1,2], that

$$m^2c^4(1-\kappa_0(r))^2 = p^2c^2 \tag{14}$$

with

$$\kappa_0(r) = \frac{2GM}{c^2 r} = 2\kappa(r) \tag{14'}$$

which means that they obey the law commensurate with the effect of light deflection in a gravitational field. In the final section we will indicate that Eq.(14) is compatible with the Jebsen-Birkoff stationary, spherically symmetric solution.

4. The Schwarzschild metric

In order to demonstrate the compatibility with the Schwarzschild metric we will introduce and generalize the following familiar formalism

$$\langle \mathbf{r}|\mathbf{p}\rangle = (2\pi\hbar)^{-3/2} e^{i/\hbar\tilde{\mathbf{r}}\cdot\mathbf{p}} \tag{15}$$

where

$$\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}; \;\; \mathbf{p} = \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} \tag{16}$$

to four dimensions. We simply obtain

$$\left\langle \mathbf{r},-ict \middle| \mathbf{p},\frac{iE}{c} \right\rangle = (2\pi\hbar)^{-2} e^{i/\hbar(\tilde{\mathbf{r}}\cdot\mathbf{p}-Et)} \tag{17}$$

using conventional notation (noting the complex conjugate in the bra-position, to subscribe to a complex symmetric construction)

$$\langle \mathbf{x}^*|\mathbf{\Pi}\rangle = (2\pi\hbar)^{-2} e^{i/\hbar(\tilde{\mathbf{x}}\cdot\mathbf{\Pi})} \tag{18}$$

with the evident definitions

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \\ ict \end{pmatrix}; \;\; \mathbf{\Pi} = \begin{pmatrix} p_x \\ p_y \\ p_y \\ iE/c \end{pmatrix} \tag{19}$$

We will now rewrite our previous equations explicitly displaying the usual operator identifications, i.e.

$$\begin{aligned} &\vec{\mathbf{p}} = -i\hbar\nabla \\ &E = i\hbar\frac{\partial}{\partial t} \\ &\vec{\mathbf{\Pi}} = -i\hbar(\nabla,\ \mathrm{i/c}\frac{\partial}{\partial t}) \end{aligned} \tag{20}$$

getting the result of the operator secular equation as (identity operator suppressed)

$$\lambda^2 = (E^2 - p^2c^2) = -c^2\vec{\mathbf{\Pi}}\cdot\vec{\mathbf{\Pi}} = -c^2\tilde{\mathbf{\Pi}}\cdot\mathbf{\Pi} = -c^2\Pi^2 = m_0^2c^4 \tag{21}$$

Incidentally we note that

$$\Pi^2 = -\hbar^2\left(\Delta - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \tag{22}$$

and

$$\left\langle \mathbf{x}^* \middle| -\Pi^2 \middle| \mathbf{\Pi} \right\rangle = m_0^2c^2\left\langle \mathbf{x}^* \middle| \mathbf{\Pi} \right\rangle \tag{23}$$

Equation (23) can be taken to define the restmass uniquely. In Ref.[1] we obtained the relation

$$m = \frac{m_0(1-\kappa(r))}{\sqrt{1-2\kappa(r)}} = \frac{\lambda_0}{\sqrt{1-2\kappa(r)}} \tag{24}$$

where λ_0 is the positive (real part) eigenvalue of the secular equation corresponding to

$$m\begin{pmatrix} (1-\kappa(r)) & \kappa(r) \\ \kappa(r) & -(1-\kappa(r)) \end{pmatrix}. \tag{25}$$

Since the eigenvalues of the matrix Eq.(25) may come out non-real in our complex symmetric setting, i.e.

$$m = m_r - i\Gamma;\ \lambda_0 = \lambda_r - i\Gamma_0 \tag{26}$$

we find by projecting out the real and complex parts of (24, 25) that

$$\tau = \tau_0\sqrt{1-2\kappa(r)};\ \Gamma = \frac{\hbar}{\tau};\ \Gamma_0 = \frac{\hbar}{\tau_0} \tag{27}$$

or for differential “times”

$$d\tau^2 = d\tau_0^2(1-2\kappa(r)) \tag{28}$$

Eqs.(14) and (28) implies that we can introduce the gravitational interaction as follows in the present case of $m_0 \to 0$ (static, spherically symmetric case), i.e.

$$\Pi^2 \to \Pi^2_{\text{grav}} = \left(1-\frac{2Gm}{c^2r}\right)^{-1} p_r^2 - \left(1-\frac{2Gm}{c^2r}\right)\frac{E^2}{c^2} \tag{29}$$

or in its appropriate symmetrised form

$$\Pi^2 \to \Pi^2_{\text{grav}} = \left(1-\frac{2Gm}{c^2r}\right)^{-1/2} p_r^2 \left(1-\frac{2Gm}{c^2r}\right)^{-1/2} - \left(1-\frac{2Gm}{c^2r}\right)^{1/2}\frac{E^2}{c^2}\left(1-\frac{2Gm}{c^2r}\right)^{1/2} \tag{29'}$$

or

$$\Pi^2 \to \Pi^2_{\text{grav}} = p_r\left(1-\frac{2Gm}{c^2r}\right)^{-1} p_r - \frac{E}{c}\left(1-\frac{2Gm}{c^2r}\right)\frac{E}{c} \tag{29''}$$

where we have made a change of the coordinate-, and the recipocal coordinate system: $\mathbf{x'} = \boldsymbol{\alpha}^{-1}\mathbf{x}$, $\mathbf{\Pi'} = \boldsymbol{\alpha}\mathbf{\Pi}$, where $\boldsymbol{\alpha}$ in general is a 4×4 similarity transformation, but here restricted to 2 dimensions in the basis (r, ict) and (p_r, iE/c) with corresponding modifications in Eqs. (15-19)

$$\boldsymbol{\alpha} = \begin{pmatrix} \left(1-\frac{2Gm}{c^2r}\right)^{-1/2} & 0 \\ 0 & \left(1-\frac{2Gm}{c^2r}\right)^{1/2} \end{pmatrix} \tag{30}$$

using only one space dimension, r. Rather than using the traditional covariant formalism we have used simple matrix algebra, see e.g. [9], to analyse the consequences of the transformation. From Eqs.(29) we obtain directly that

$$\begin{gathered} \left\langle (r,\ ict)^* \middle| \Pi^2 \middle| p_r,\ iE/c \right\rangle \to \left\langle (r',\ ict')^* \middle| \Pi'^2 \middle| p'_r,\ iE'/c \right\rangle = \\ \left\langle (r',\ ict')^* \middle| \Pi^2_{\text{grav}} \middle| p'_r,\ iE'/c \right\rangle = \left\langle (r,\ ict)^* \middle| \Pi^2_{\text{grav}} \middle| p_r,\ iE/c \right\rangle \end{gathered} \tag{31}$$

Thus we have succeeded to build the gravitational interaction into the formulation such that the surrounding field here only appears as an effect of the geometry characterizing the gravitational source. Hence a coordinate transformation α in general defines appropriate boundary conditions for the quantum formulation. From Eqs. (14) and (28) follows also a more direct non-quantum argument of the compatibility of the Schwarzschild metric in the spherically symmetric, static vacuum.

It is, as a result of the present study, interesting to note that the present formulation of gravitational interactions is essentially "classical" or rather quasi-classical outside the domain boundary characterized by the Schwarzschild radius. The contrast is that the interpretation of the laws of relativity stems from the viewpoint of the quantum mechanical superposition principle. Yet, except from the occurrence of general Jordan block like singularities, the equations appear mostly to be of classical-orthodox character. The formulation inside the singularity, which is rejected by classical theories, follows from a specific degeneracy condition, see Eqs.(12, 13), where all interactions/correlations condense or unify according to Yang's ODLRO[10], see also [2] for more details.

5. Acknowledgements

The author is indebted to Prof. Jésus Vigo-Aguiar for generous support and hospitality. Supported in part by the Swedish Foundation for Strategic Research.

6. References

[1] E. BRÄNDAS, *Are Einstein's Laws of Relativity a Quantum Effect?* in Frontiers in Quantum Systems in Chemistry and Physics, eds. by J. Maruani et. al, Kluwer Academic Publishers, **Vol. 18,** (in press) 2008.

[2] E. BRÄNDAS, *Quantum Mechanics and the Special- and General Theory of Relativity*, Adv. Quant. Chem. **54**, (2008) 115-132.

[3] E. Brändas: *Some Theoretical Problems in Chemistry and Physics*. Int. J. Quant. Chem. **106**, (2006) 2836-2839.

[4] E. BALSLEV, J. M. COMBES, *Spectral properties of many-body Schrödinger operators with dilatation-analytic interactions*, Commun. Math. Phys. **22** (1971) 280-294.

[5] E. J. Brändas: *Resonances and Dilation Analyticity in Liouville Space*. Adv. Chem. Phys. **99,** (1997) 211-244.

[6] A. BOHM, M. GADELLA, M. LOEVE, S. MAXON, P. PATULEANU, C. PUNTMANN, *Gamow-Jordan Vectors and Nonreducible Density Operators from Higher Order S-Matrix Poles*, J. Math. Phys. **38** (1997) 6072-6100.

[7] J. T. Jebsen, Ark. Mat. Ast. Fys. **15,** (1921) 1-9.

[8] G. D. Birkoff, *Relativity and Modern Physics,* Cambridge University Press, 1921.

[9] P. O. LÖWDIN, *Linear Algebra for Quantum Theory*, Wiley, New York, 1998.

[10]C. N. Yang, Rev. Mod. Phys. **34,** (1962) 694-704.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Implementing a relational system for order of magnitude reasoning*

A. Burrieza[1], **A. Mora**[2], **M. Ojeda-Aciego**[2] **and E. Orłowska**[3]

[1] *Dept. Filosofía. , Univ. Málaga, (Spain)*

[2] *Dept. Matemática Aplicada., Univ. Málaga, (Spain)*

[3] *National Institute of Telecommunications, Warsaw, (Poland)*

emails: burrieza@uma.es, amora@ctima.uma.es, aciego@ctima.uma.es, orlowska@itl.waw.pl

Abstract

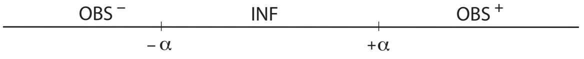
This work concentrates on the automated deduction of logics of order-of-magnitude reasoning. Specifically, a Prolog implementation is presented for the Rasiowa-Sikorski proof system associated to the relational translation $Re(OM)$ of the multimodal logic of qualitative order-of-magnitude reasoning OM.

Key words: Relational theorem proving, Rasiowa-Sikorski procedure.

1 Introduction

This paper concentrates on the logic approach to order-of-magnitude qualitative reasoning firstly introduced in [1], and further developed in [2]. Roughly speaking, the approach is based on a system with two landmarks, $-\alpha$ and $+\alpha$, which is both simple enough to keep under control the complexity of the system and rich enough so as to permit the representation of a subset of the usual language of qualitative order-of-magnitude reasoning.

The intuitive representation of the underlying frames is given in the picture below, where $-\alpha$ and $+\alpha$ represent respectively the greatest negative observable and the least positive observable, partitioning the real line in classes of positive observable $\textsc{Obs}^+$, negative observable $\textsc{Obs}^-$ and non-observable numbers $\textsc{Inf}$:



*Partially supported by TIC06-15455-C03-01 and P06-FQM-02049

In [3], the paradigm 'formulas are relations' formulated in [9] was applied to the modal logic for order-of-magnitude (OM) reasoning introduced in [2], obtaining a relational logic $Re(OM)$ based on algebras of relations generated by some relations specific to the frames of OM-logics; after a translation from the language of OM-logics to the language of $Re(OM)$, a deduction system for $Re(OM)$ in the Rasiowa-Sikorski style [10] was presented, paving the way for applicative research on the implementation of the proof procedure. The main contribution of this work consists in the development of a Prolog implementation of the Rasiowa-Sikorski proof procedure introduced in [3]; it is worth to note that our proof system is modular, in that adding new semantic constrainsts to the logic implies adding new deduction rules or axiomatic sets, and not implementing a new system from scratch.

The structure of the paper is the following: in Section 2, the language $Re(OM)$ is introduced, together with the relational proof system; then, the main contribution of the paper, is presented in Section 3, which contains the implementation in Prolog of the relational procedure in Section 4, some executions of the Prolog engine are shown, the input formulas are taken from the axiom system for the logic as presented in [2]; Section 5, finally, concludes and presents prospects for future work.

2 The language $Re(OM)$: the RS proof system

As stated in the introduction, the language OM was translated in [3] into a relational one in order to take benefit from the RS proof procedure. As usual, the main idea of the relational formalisation is to interpret formulas of nonclassical logics as relations which are the elements of algebras of relations from a suitable class. For limitation of the length of the paper, we reveal only the language $Re(OM)$.

We recall here the definition of The syntax of $Re(OM)$ [1].

The alphabet of $Re(OM)$ consists of the disjoint sets listed below:

- A (nonempty) set $\mathbb{OV} = \{x, y, z, \dots\}$ of object variables.
- A set $\mathbb{OC} = \{\alpha^-, \alpha^+\}$ of object constants.
- A (nonempty) set $\mathbb{RV} = \{P, Q, R, \dots\}$ of binary relation variables.
- A set $\mathbb{RC} = \{1, 1', \aleph^-, \aleph^+, <, \sqsubset, \prec\}$ of relation constants denoting, respectively, the universal relation, the identity relation, the constant relations for $-\alpha$ and $+\alpha$, and the three ordering relations related to the three modalities of the OM language.
- A set $\mathbb{OP} = \{-, \cup, \cap, ;, ^{-1}\}$ of relational operation symbols which are interpreted as the opposite, the union, the intersection, the composition and the inverse of a relation.

Now, the set of relation terms and formulas of $Re(OM)$ is given as follows:

[1] We show the syntax and for more details the reader is suggested to consult [3].

- The set of *relation terms* $\mathbb{RT}$ is the smallest set of expressions that includes all the relational variables and relational constants and is closed with respect to the operation symbols from $\mathbb{OP}$.
- The set $\mathbb{FR}$ of *formulas*, consists of expressions of the form xRy where x, y denote individual (or object) variables or constants and R is a relational term built from the relational variables and the relational operators.

We will now concentrate on the presentation of the RS proof system for $Re(OM)$. Let us recall that, given a relational formula xAy, where A may be a compound relational expression, we successively apply *decomposition or specific rules*. In this way a tree is formed whose root consists of xAy and each node (except the root) is obtained by an application of a rule to its predecessor node. The application of rules is stopped on a node when an *axiomatic set* (which denotes a tautological formula) has been obtained, or when none of the rules is applicable to the formulas in this node. Such a tree is referred to as a *proof tree* for the formula xAy. A branch of a proof tree is said to be *closed* whenever it contains a node with an axiomatic set of formulas. A tree is closed iff all of its branches are closed.

Our system considers the usual rules for the calculus of binary relations with equality (these rules are not shown explicitly here due to length restrictions, see for instance [6]). New specific rules are included in order to handle the specific object and relation constants of the language $Re(OM)$. These rules are shown in Fig. 1, in which the new variables occurring in the denominator of some rules denote *any* variable occurring in the branch.

The *axiomatic sets* of $Re(OM)$ shown below state valid formulas of the system which allows for stopping the procedure on a given branch.

$$\{x1y\} \qquad \{x1'x\} \qquad \{x{-}Ry, xRy\} \qquad \{\alpha^- < \alpha^+\}$$

where $x, y \in \mathbb{OS}$ and $R \in \mathbb{RT}$.

3 Prolog implementation of the relational system

In this section, we introduce the Prolog implementation[2] of the relational system given above.

Once the system receives as input the relational formula to be checked, it generates a proof tree, whose leaves contain sets of relational terms to be proved. The input formula gets proved when Prolog closes all the leaves in the proof tree.

To begin with, the relations have to be encoded as predicates. This is done as follows: A relational formula xRy, where x, y are object variables and R is a relational term represented as the Prolog fact:

$$rel(address, R, x, y)$$

[2] The full implementation (developed in SWI-Prolog Version 5.6.33 for Windows platform) is available from the address `http://homepage.mac.com/alicauchy/`.

$$\frac{x\aleph^{-}y}{x1'\alpha^{-}, x\aleph^{-}y}\ \textbf{(c1a)}\quad \frac{x{-}\aleph^{-}y}{x{-}1'\alpha^{-}, x{-}\aleph^{-}y}\ \textbf{(c1b)}\quad \frac{x\aleph^{+}y}{x1'\alpha^{+}, x\aleph^{+}y}\ \textbf{(c2a)}\quad \frac{x{-}\aleph^{+}y}{x{-}1'\alpha^{+}, x{-}\aleph^{+}y}\ \textbf{(c2b)}$$

$$\frac{x<\alpha^{+}}{x1'\alpha^{-}, x<\alpha^{+}}\ \textbf{(c3)}\quad \frac{x{-}\sqsubset y}{x1'\alpha^{-}, x{-}\sqsubset y}\ \textbf{(c4)}\quad \frac{x{-}\sqsubset y}{y1'\alpha^{+}, x{-}\sqsubset y}\ \textbf{(c5)}$$

$$\frac{x\leq\alpha^{-}, \alpha^{+}\leq x, x-\sqsubset y}{x\leq\alpha^{-}, \alpha^{+}\leq x, x-\sqsubset y, y\leq\alpha^{-}}\ \textbf{(c6)}\qquad \frac{x\leq\alpha^{-}, \alpha^{+}\leq x, x-\sqsubset y}{x\leq\alpha^{-}, \alpha^{+}\leq x, x-\sqsubset y, \alpha^{+}\leq y}\ \textbf{(c7)}$$

$$\frac{\alpha^{-}\leq x, x-\sqsubset y}{\alpha^{-}\leq x, x-\sqsubset y, \alpha^{-}<y}\ \textbf{(c8)}\quad \frac{x-<y, \alpha^{-}<y}{x-<y, \alpha^{-}<y, x-\sqsubset y}\ \textbf{(c9)}\quad \frac{x-<y, x<\alpha^{+}}{x-<y, x<\alpha^{+}, x-\sqsubset y}\ \textbf{(c10)}$$

$$\frac{x\leq\alpha^{-}, \alpha^{+}\leq x, y\leq\alpha^{-}, \alpha^{+}\leq y, x\sqsubset y}{x\leq\alpha^{-}, \alpha^{+}\leq x, y\leq\alpha^{-}, \alpha^{+}\leq y, x\sqsubset y, x<y}\ \textbf{(c11)}\qquad \frac{x{-}\sqsubset y}{x{-}\sqsubset y, x{-}{<}y}\ \textbf{(c12)}$$

$$\frac{}{x<x}\ \textbf{(Iref)}\quad \frac{}{y{-}{<}x \mid x{-}{<}y \mid x{-}1'y}\ \textbf{(Lin)}\quad \frac{}{x\sqsubset y \mid x-\sqsubset y}\ \textbf{(cut-}\ \sqsubset\textbf{)}\quad \frac{xRy}{xRy, xRz, \mid xRy, zRy}\ \textbf{(Tran)}$$

$$\frac{x<y}{x\prec y, x<y}\ \textbf{(n-0)}\quad \frac{x\prec z}{x\prec y, x\prec z \mid y<z, x\prec z}\ \textbf{(n-i)}\quad \frac{x\prec z}{x<y, x\prec z \mid y\prec z, x\prec z}\ \textbf{(n-ii)}$$

$$\frac{\alpha^{+}\leq y}{\alpha^{-}<x, \alpha^{+}\leq y \mid x<\alpha^{+}, \alpha^{+}\leq y \mid x\prec y, \alpha^{+}\leq y}\ \textbf{(n-iii)}$$

$$\frac{y\leq\alpha^{-}}{\alpha^{-}<x, y\leq\alpha^{-} \mid x<\alpha^{+}, y\leq\alpha^{-} \mid y\prec x, y\leq\alpha^{-}}\ \textbf{(n-iv)}$$

Figure 1: Specific rules for $Re(OM)$

The first argument contains a list of integers which define the position of the node in the proof tree, as it has been generated during the proof process.

Example 1 *The formulas contained in a leaf of a proof tree are read disjunctively, hence an expression as* $xRy \cup xSy \cup x\aleph^{-}y \cup x(\sqsubset;(a;1)^{-})^{-}y$ *is translated into the following four facts in Prolog:*[3]

```
rel([1],r,x,y).
rel([1],opp(alephm),x,y).
rel([1],s,x,y).
rel([1],opp(comp(sqsub,opp(comp(a,univ)))),x,y).
```

The (addresses of the) open leaves are stored in a list, which is handled by the predicate `open_leaves`. For instance, the predicate `open_leaves([n])` states that it is necessary to prove the validity of the set of relations stored in node `[n]`.

As expected, the initial relational terms are valid if and only if all the leaves in the tree can be closed.

Expressing axiomatic sets and rules

When Prolog detects a relation representing an *axiomatic set*, the corresponding leaf is deleted and the user informed by means of the `remove_leaf` predicate. For instance,

[3] As Prolog only manipulates text, some symbols are renamed accordingly to its reading. For instance, $\aleph^{-}$ is translated into `alephm`; the composition operator `;` is translated into *comp*, the operator $\sqsubset$ is translated into `sqsub`, etc.

if either $x1'x$ (`rel(Leaf,equal,X,X)`) occurs in the set of relations of the leaf `Leaf`, it is removed because of the occurrence of an axiomatic set.

```
axiomatic_set:- rel(Leaf,equal,X,X),
                  remove_leaf(Leaf,[rel(Leaf,equal,X,X)]),!.

axiomatic_set:- rel(Leaf,univ,X,Y),
                remove_leaf(Leaf,[rel(Leaf,univ,X,Y)]),!.

axiomatic_set:- rel(Leaf,<,alpham,alphap),
                remove_leaf(Leaf,[rel(Leaf,univ,X,Y)]),!.
```

A *rule* in $Re(OM)$ has the following general form: $\frac{\Phi}{\Phi_1|\dots|\Phi_n}$ where $\Phi_1,\dots,\Phi_n$ are non-empty sets of formulas and Φ is a finite (possibly empty) set of formulas.

The application of a rule like the previous one to a leaf assumes it is labelled by a set X of formulas satisfying $\Phi \subseteq X$, then the leaf branches into n new branches, each one with the set of formulas $(X \setminus \Phi) \cup \Phi_i, i = 1\dots,n$.

In general, due to the particular nature of the rules of $Re(OM)$, whenever a rule is applicable, it can be applied again on the resulting leaves, but this kind of behaviour is obviously undesirable. In order to avoid repeated applications of rules against the same formulas each application of a rule is stored in a list.

The implementation of a rule can be roughly stated as follows: firstly, the preconditions (contained in the numerator of the rule) are checked, in order to know whether the rule is applicable; if affirmative, and provided that the rule has not been previously applied against the same arguments, the rule is displayed on the screen and stored as used; finally, the leaf is branched and new labels are attached to each new leaf as stated above.

In order to obtain a rough idea of how a rule is encoded, let us consider the standard rule for the union of relations $\dfrac{x(R\cup S)y}{xRy, xSy}$ (**uni**), its encoding is:

```
uni(Leaf):- rel(Leaf,uni(R,S),X,Y),
            new_deduced_rels([rel(Leaf,R,X,Y),rel(Leaf,S,X,Y)]),
            \+rule_used(Leaf,uni,[rel(uni(R,S),X,Y)]),
            write_rule('Union', [rel(Leaf,uni(R,S),X,Y)],
                                [rel(Leaf,R,X,Y), rel(Leaf,S,X,Y)]),
            update_leaf([rel(Leaf,R,X,Y),rel(Leaf,S,X,Y)]).
```

In order to start explaining the most interesting features of the implementation, let ys consider a specific (non-standard) rule (n-i) below:

$$\frac{x \prec z}{x \prec y, x \prec z \mid y < z, x \prec z} \text{ (n-i)} \quad y \text{ any variable}$$

This rule is implemented by using the following code:

```
ni(Leaf):- rel(Leaf,prec,X,Z),
        new_deduced_rels([rel(Leaf,prec,X,Y), rel(Leaf,<,Y,Z)]),
        \+rule_used(Leaf,ni,[rel(prec,X,Z)]),
        any_variable('ni  (prec) ',Leaf,[rel(Leaf,prec,X,Z)],Y),!,
        write_rule('ni (prec) ', [ rel(Leaf,prec,X,Z)],
                          [ rel(Leaf,prec,X,Y),rel(Leaf,<,Y,Z)]),
        branch(Leaf,2),
        update_leaf(Leaf,2,[[rel(Leaf,prec,X,Y)]
                                ,[rel(Leaf,<,Y,Z)]]),!.
```

In the three first lines, the rule checks that $x \prec z$ is in the set of relations, that the relations introduced by the rule are new (`new_deduced_rels`) and that the rule has not been previously applied (`rule_used`). Then, note that, as stated in the rule, the variable y in the denominator has to be any of the variables or constant object occurring in the branch (this situation is similar to that of the free tableaux systems, in which the γ rule instantiates a variable by any of the constants occurring in the branch, whereas the δ rule always introduces a new constant). The predicate `any_variable` chooses some constant or variable occurring in the branch (an optimized version of this task is given in Section 3). The predicate `branch(Leaf,2)` branches the current leaf into two new leaves, and copies all the formulas of the current leaf to the two new leaves. The predicate `copyToLeaves` appends $x \prec y$ to the first leaf and $y < z$ to the second leaf.

The proof procedure

The implementation of a full and automated proof procedure is roughly sketched here. The inference engine examines the first leaf of the tree that the proof system needs to check and tries to apply the rules to the relations containing this leaf. As stated previously, the predicate `open_leaves` stores the leaves which has not been closed so far. The inference engine tries to apply some rule to the given leaf, while the tree has open leaves.

The order in which the engine tries to apply the rules is crucial. Clearly, the rules which do not generate new branches are at the beginning; among these rules we have some primitive rules (either standard or specific), then some selected derived rules have been implemented directly as primitive, in order to avoid excessively long proofs. Finally, the system tries to apply the rules that generate new branches.

Whenever a non-closed leaf does not admit any of the rules in the list, then the system asks the user about considering some cut-like rule (a rule without relations in the numerator).

After an application of the procedure, and provided that a closed tree has been obtained, the system provides a list of the rules used in the proof; this is done by the predicate `table_of_used_rules`. As an example, consider the output obtained from the following relational formula (which corresponds to the Axiom c4 of the system for the logic OM, the formula $\alpha^- \to \overrightarrow{\blacksquare} A$, see [2]):

```
rel([1], uni(opp(alephm),opp(comp(sqsub,opp(comp(p,univ))))),x,y).
```

The system traces, in reverse ordering, the rules applied in order to close the tree for the input term:

```
 OK. No more open leaves. VALID.
table_of_used_rules([1], c5, [rel(opp(sqsub), z, x)]).
table_of_used_rules([1], c4, [rel(opp(sqsub), z, x)]).
table_of_used_rules([1], c2b, [rel(opp(alephp), x, y)]).
table_of_used_rules([1], notinverse,[rel(opp(inv(sqsub)), x, z)]).
table_of_used_rules([1], not2, [rel(opp(opp(comp(p, univ))), z, y)]).
table_of_used_rules([1], notcomp,[rel(opp(comp(inv(sqsub), opp(comp(p, univ)))), x, y)]).
table_of_used_rules([1], uni, [rel(uni(opp(alephp), opp(comp(inv(sqsub),
                                                      opp(comp(p, univ))))), x, y)]).
```

Phantom variables: postponing the choice

There are several rules in the relational system for $Re(OM)$ which exhibit the same behavior that Rule (n-i) regarding the new variables introduced. We saw that the rule branches the leaf into two new leaves, and appends $x \prec y$ to the first leaf and $y < z$ to the second leaf, where y is "*any variable*" occurring in the branch. In principle, we have as many different instantiations of the rule as values can be chosen for y. If we do not take this into account, the proof tree might grow in an uncontrolled manner.

We introduce a non-instantiated variable (so-called "phantom variable") and delay its actual instantiation until we have some guarantees, by a unification process, that it will generate axiomatic sets. Thus, a *phantom variable* is a special case of variable whose possible instantiations are constrained to belong to the set of variables or constants occurring in the leaf.

The use of phantom variables is crucial for an adequate performance of the implementation, although it initially implied the need to rewrite the code for the axiomatic sets in order to make them parameterized. For instance, recall that if the axiomatic set $\alpha^- < \alpha^+$ is present in a leaf, then the leaf will be closed; as a result, $X < \alpha^+$ will be an axiomatic set provided that X is a phantom variable which can be instantiated by α^-.

4 Experimental results and examples

As the relational proof procedure was proved to be complete in [3], the first choice of formulas to prove with the implementation has been the set of axioms of the system given in [2]. The implementation has been tested against all the axioms in the system[4] with the result that every axiom has been automatically proved. This is an important matter, since so far no result about the decidability of $Re(OM)$ has been obtained.

In this section we comment in detail the performance of the implementation on the relational translation of two specific axioms of OM.

Example 2 *Let us consider the formula $\alpha^- \to \overrightarrow{\blacksquare} A$, corresponding to Axiom c4 from [2]. Its relational translation is*

$$x(-\aleph^- \cup -(\sqsubset; -(A;1)))y$$

which, in turn, is translated into Prolog as:

[4] The full trace of execution of the procedure applied on all the axioms of [2] can be obtained from the address `http://homepage.mac.com/alicauchy/`.

```
rel(1,opp(alephm),x,y).
rel(1,opp(comp(sqsub,opp(comp(a,univ)))),x,y).
```

Now, the program is called to satisfy the predicate:

?engine(’reomAxiomc4.pl’,’logc4.txt’).

The following report in logc4.txt file is returned:

```
------>Input file: reomAxiomc4.pl
THE ENGINE IS RUNNING
--->opp composition Rule
[rel(1, opp(comp(sqsub, opp(comp(a, univ)))), x,y)]
---------------------------------------------------------------
[rel(1, opp(sqsub), x, z), rel(1,opp(opp(comp(a, univ))), z, y)]

---->c1b (opp aleph-)  Rule
[rel(1, alephm, x, y)]
------------------------------------------------
[rel(1, equal, x, alpham), rel(1, alephm, x, y)]

---->c4 (notsqsubset)  Rule
[rel(1, opp(sqsub), x, z)]
----------------------------------------------------
[rel(1, equal, x, alpham), rel(1, opp(sqsub), x,z)]

Found axiomatic set. Branch: 1
    - Axiomatic set: [rel(1, opp(equal), x, alpham),
                      rel(1, equal, x, alpham)]
    - Deleted relations in branch 1
 OK. No more open leaves.
```

□

The following example is more complete than the previous one, as it branches the proof tree and, in addition, uses phantom variables.

Example 3 *Let us consider the formula* $\overleftarrow{\Diamond}\alpha^- \vee \alpha^- \vee \overrightarrow{\Diamond}\alpha^-$*, corresponding to Axiom c1 from [2]. Its relational translation is*

$$x((>;\aleph^+)\cup\aleph^+\cup(<;\aleph^+))y$$

which in Prolog has the following form:

```
rel([1], comp(>, alephp), x, y).
rel([1], alephp, x, y).
rel([1], comp(<, alephp), x, y).
```

Now, the program is called to satisfy the predicate:

?engine(’reomAxiomc1.pl’,’logc1.txt’).

After applying some rules, the system detects the possibility of using one phantom variable, the following information is displayed on the screen:

```
We can apply the following rules:
---->comp Rule
[rel([1], comp(>, alephp), x, y)]
-----------------------------------------------------------
rel(new_leaf1, >, x, var) | rel(new_leaf2, alephp, var, y)

where var can be either:
   - any variable from: [x, y]
   - or alpham or  alphap.
We can use a non-instantiated variable (phantom).
Introduce the desired var or 0 for phantom variable.
```

Now, the user can either introduce any of the possible values, or let the system introduce a phantom variable. In this example, the system is always said to introduce phantom variables (which are denoted as t1*,* t2*, etc). Thus, the log file of this example continues as follows:*

```
|: 0
---->comp Rule
[rel([1], comp(>, alephp), x, y)]
----------------------------------------------------
rel([1, 1], >, x, t1) | rel([1, 2], alephp, t1, y)
```

The system continues applying rules automatically until a new composition (comp) rule is applied. Note that, in the leaf (1,1,2) we would obtain an axiomatic set if t2 is substituted by alphap.

```
---->comp Rule
[rel([1, 1], comp(<, alephp), x, y)]
----------------------------------------------------------
rel([1, 1, 1], <, x, t2) | rel([1, 1, 2], alephp, t2, y)

Substitute in all relations variable phantom:t2 by alphap
```

This instantiation provides an extra piece of information which allows eventually to close all the open branches of the proof tree. More details can be seen in the demos available in the web. □

5 Conclusions and future work

We have presented a first implementation in Prolog of the relational proof system for the logic of qualitative order-of-magnitude reasoning. The system has been tested against the axiom system provided in [2], and all the axioms of the system have been automatically proved. This is an important matter, since so far no result about the decidability of $Re(OM)$ has been obtained.

As future work, the implementation will be improved in several directions. On the one hand, we want to add more interaction with the user during the proof process. When the system does not close the proof tree, some cut-like rule might be needed and the user should be asked to provide some clue on this although, in some situations, it is possible for the system to suggest the use of some of these rules. On the other hand, the graphical aspect of the interface should be enhanced, allowing the user to specify directly the requirements by using the standard OM logic, which is more intuitive than its relational translation into $Re(OM)$.

References

[1] A. Burrieza and M. Ojeda-Aciego. A multimodal logic approach to order of magnitude qualitative reasoning. *Lect. Notes in Artificial Intelligence*, 3040:431–440, 2004.

[2] A. Burrieza and M. Ojeda-Aciego. A multimodal logic approach to order of magnitude qualitative reasoning with comparability and negligibility relations. *Fundamenta Informaticae*, 68:21–46, 2005.

[3] A. Burrieza, M. Ojeda-Aciego, and E. Orłowska. Relational approach to order-of-magnitude reasoning. *Lect. Notes in Computer Science*, 4342:105-124, 2006.

[4] J. Dallien and W. MacCaull. RelDT—a dual tableaux system for relational logics, 2005. Available from `http://logic.stfx.ca/reldt/`

[5] A.Formisano, E. Omodeo, and E. Orłowska. A PROLOG tool for relational translation of modal logics: A front-end for relational proof systems. In: B. Beckert (ed) TABLEAUX 2005 Position Papers and Tutorial Descriptions. Universität Koblenz-Landau, Fachberichte Informatik No 12, 2005, 1-10. System available from `http://www.di.univaq.it/TARSKI/transIt/`

[6] J. Golińska-Pilarek and E. Orłowska. Tableaux and dual tableaux: Transformation of proofs. *Studia Logica* 85(3):283–302, 2007.

[7] B. Konikowska. Rasiowa-Sikorski deduction systems in computer science applications. *Theoretical Computer Science* 286:323–366, 2002

[8] E. Orłowska. Relational interpretation of modal logics. In H. Andreka, D. Monk, and I. Nemeti, editors, *Algebraic Logic*, volume 54 of *Colloquia Mathematica Societatis Janos Bolyai*, pages 443–471. North Holland, 1988.

[9] E. Orłowska. Relational semantics for nonclassical logics: Formulas are relations. In J. Wolenski, editor, *Philosophical Logic in Poland*, page 167–186. Kluwer, 1994.

[10] H. Rasiowa and R. Sikorski. *Mathematics of Metamathematics.* Polish Scientific Publishers, 1963.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Estimating topological entropy from individual orbits

Jose S. Cánovas[1]

[1] *Department of Applied Mathematics and Statistics, Technical University of Cartagena*

emails: `jose.canovas@upct.es`

Abstract

A method for computing topological entropy from individual orbits is given.

Key words: Topological entropy, interval maps, permutations
MSC 2000: AMS 37E05)

Given a continuous interval map $f : [0,1] \to [0,1]$, its topological entropy $h(f)$ (see [1]) is an useful tool to decide whether the dynamical behavior of f is complicated. Roughly speaking, it measures the different number of possible orbits of f, and hence, its computation needs the evaluation of several orbits.

Let S_n be the set of permutations of length n and consider a trajectory $T(x) = (f^m(x))_{m=0}^{\infty}$ of f, $x \in [0,1]$. Let p_n be the number of permutations $\pi \in S_n$ for which there is k such that $f^{k+\pi(1)}(x) < f^{k+\pi(2)}(x) < ... < f^{k+\pi(n)}(x)$. Define the topological entropy of the trajectory as

$$h(T(x)) = \limsup_{n\to\infty} \frac{1}{n} \log p_n.$$

It is simple to prove that the topological entropy of a periodic trajectory is zero. Our main result is that if the map f is piecewise monotone and transitive, then there is a point x such that $h(f) = h(T(x))$. We use this fact to give some numerical approximations to the topological entropy of a continuous interval map.

References

[1] R. Bowen, *Entropy for group endomorphism and homogeneous spaces*, Trans. Amer. Math. Soc. **153** (1971), 401–414.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Switched linear systems: A study using local automata

Miguel V. Carriegos[1], **H. Diez-Machío**[2], **A. Prieto-Castro**[2] **and L. Rivas-Morán**[2]

[1] *Departamento de Matemáticas, Universidad de León. SPAIN*

[2] *Instituto Nacional de Tecnologías de la Comunicación (INTECO), SPAIN*

emails: `miguel.carriegos@unileon.es`, `hector.diez@inteco.es`,
`alejandro.prieto@inteco.es`, `lorena.rivas@inteco.es`

Abstract

Regular switched linear systems are introduced as a generalization of switched linear systems. Reachability properties are studied by using localization of regular languages.

Key words: hybrid system, reachability, equivalence
MSC 2000: 93B03, 93B05, 93B11, 68Q45

1 Introduction

Hybrid systems have been attracting much attention in the recent past years because of the arising problems are not only academically challenging but also of practical importance in a wide field of applications ranging from manufacturing systems to information processes and modeling ecosystems, among others [7].

Switched linear systems belong to a special class of hybrid control systems which comprises a collection of subsystems described by linear dynamics (differential/difference equations) together with a switching rule that specifies the switching between the subsystems (see [2], [3] and [7]-[18]).

In this paper we consider *programmable switched systems*; that is, sequential switched linear systems

$$\Gamma : \underline{x}(t+1) = A_{\sigma(t)}\underline{x}(t) + B_{\sigma(t)}\underline{u}(t)$$

where the switching signals $\sigma(0)\sigma(1)\sigma(2)... \in \Sigma^*$ belong to a formal language $L_\Gamma \subseteq \Sigma^*$ of admissible sequences of commands of system Γ. First we restrict to the case of $L_\Gamma \subseteq \Sigma^*$ being regular. This is actually equivalent to saying that switching signals are governed by a finite automaton.

We introduce a method in order to describe a regular switched linear system Γ in terms of a finite automaton for the regular language L_Γ of admissible switching signals.

Note that our definition will be natural generalization of the sequential case of both linear systems and switched linear systems.

First we recall the definition of switched linear system from [11] in order to focus the object of our study.

Definition 1.1 *A switched linear system is given by*

$$\begin{aligned} \delta x\left(t\right) &= A_{\sigma(t)} x\left(t\right) + B_{\sigma(t)} u\left(t\right) y\left(t\right) \\ y\left(t\right) &= C_{\sigma(t)} x\left(t\right) \end{aligned}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control input, $y \in \mathbb{R}^p$ is the output, σ is the piecewise constant switching signal taking value from the finite index set $\mathcal{I} = \{0, ..., s\}$. For unified presentation, δ denotes the derivative operator $\frac{d}{dt}$ in the continuous case and the forward shift operator $\delta x(t) = x(t+1)$ in the sequential case.

This paper deals with the study of reachability properties, hence output maps play no role. Consequently we will not consider any output equation (or, equivalently, consider that $x = y$ and $C_\sigma = \mathbf{1} = (\delta_{ij})$ is the identity matrix).

The evolution of the discrete signals $\sigma(t)$ can be described in a variety of ways. In Switched Linear Systems, σ is an unknown, deterministic and finite-valued input freely chosen by the controller, in Jump-Markov Linear Systems σ is a Markov Chain governed by the transition probabilities $\pi(i,j) = p(\sigma(t+1) = j/\sigma(t) = i)$, while in Piecewise Affine Linear Systems $\sigma(-)$ is piecewise affine function of the internal states.

In this paper we deal with the case of switching signals are restricted to a regular language (of admissible sequences of commands) $L \subseteq \Sigma^*$. This formalization is inspired by notion of a programmable system, where not every sequence of switching signals are allowed. This is a generalization of the case of Switched Linear Systems because in the latter case $L = \Sigma^*$ is a particular regular language (every sequence of commands is admissible).

On the other hand, this is a more general framework that is closer to the topic of hybrid linear systems (see [10] and [14]).

The paper is organized as follows: Section 2 is devoted to review switched linear systems by giving some motivating examples and main results about behavior of a switched linear system. We also study reachability property by giving a new method which reduces the case (A_σ, B_σ) to the case (A_σ, B).

Section 3 deals with regular switched linear systems which generalize switched linear case. Here we only allow a subset (formal language) of admissible switching signals. We focus reachability property and point out some obstructions. These obstructions are attacked in Section 4 by reducing to the case of local languages and providing an Algorithm to check reachability of a local switched linear system.

2 Switched Systems

Definition 2.1 *A Switched Linear System is given by an evolution equation on the form*

$$x(t+1) = A_{\sigma(t)}x(t) + B_{\sigma(t)}u(t)$$

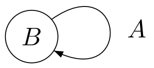
where $x \in \mathbb{K}^n$ is the internal state of system, $u \in \mathbb{K}^m$ is a external input (or control) of system, together with

$$\sigma(t) = \varphi(t, \sigma(t-1), x(t))$$

which is the next command function. Commands (or switches) are finite sequences (words) on the finite alphabet $\Sigma = \{0, ..., s-1\}$.

Switching rules can be described via a digraph. Next we write down explicit digraphs for some interesting examples. The nodes are, in some sense, the components or subsystems of the switched system.

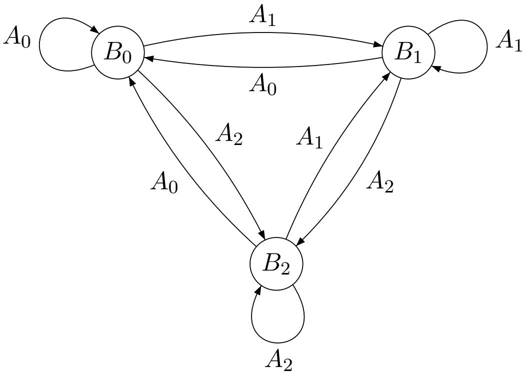
Example 2.2 *The following automaton represents a classic linear system (A, B): The label B into the node represent that one is allowed to add a vector in the space (of controls) spanned by matrix B at any time, while loop labeled by A represent that the next state is obtained by applying the operator A. Hence, once a initial state $x(0)$ is fixed, we obtain the dynamics by following the loop and adding a control $Bu(t)$ at time t: $\underline{x}(t+1) = A\underline{x}(t) + B\underline{u}(t)$*



Obviously in this classical case, $\mathcal{I} = \{0\}$ is a singleton.

The general case of switched linear systems, every function $\sigma(t) = \varphi(t)$ can be selected and every switched rule can be used (see [13]). This situation is remarked in the following example.

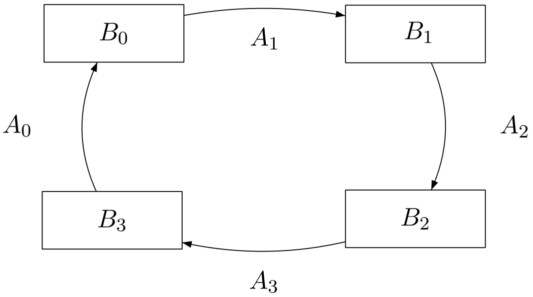
Example 2.3 *A general switched linear system with three subsystems where the controlled is allowed to change among subsystems at any time is now represented by the following full 3-state automaton*



This automaton represents a switched linear system (A_σ, B_σ) in the sense of, for example [15]. Here $\sigma \in \Sigma$ being $\Sigma = \{0,1,2\}$ the alphabet of three possible commands.

Some hybrid systems don't allow any sequence of switching between subsystems. These systems are studied in some detail in Section 3. Now we give an example of that kind of hybrid systems

Example 2.4 *Suppose that some linear systems are available to describe some ecosystem depending on the season. Fixing indices 0=Spring, 1=Summer, 2=Fall, 3=Winter; linear systems are thus denoted by (A_i, B_i) and the switching function is $\sigma(i) = (i+1)(\text{mod})4$ or, in other words we have the digraph*



It is interesting to study the behavior of a given switched linear system for a fixed sequence $\underline{\sigma}$ of commands (or switching signals) and a fixed sequence $\underline{u}$ of external inputs. First we need a preparatory result.

Lemma 2.5 *The behavior of a switched linear system Γ is given by the equalities*

$$\begin{array}{rcl}\Phi_\Gamma(x_0, \underline{\sigma}\tau, \underline{u}v) & = & A_\tau \Phi_\Gamma(x_0, \underline{\sigma}, \underline{u}) + B_\tau v \\ \Phi_\Gamma(x_0, \tau, v) & = & A_\tau x_0 + B_\tau v\end{array}$$

Where $\underline{\sigma} \in \Sigma^*$, $\tau \in \Sigma$, $\underline{u} \in (\mathbb{K}^m)^*$ and $\tau \in \mathbb{K}$.

Proof.- Direct application of the definition of switched linear system □

Theorem 2.6 *Let $\Gamma : x(t+1) = A_{\sigma(t)}x(t) + B_{\sigma(t)}u(t)$ be a switched linear system .The behavior of system Γ from initial state x_0, with sequence $\underline{\sigma} = \sigma(0)\sigma(1)\cdots\sigma(s)$ of commands, and sequence $\underline{u} = u(0)u(1),...,u(s)$ of controls is*

$$\Phi_\Gamma(x_0, \underline{\sigma}, \underline{u}) = A_{\sigma(s)}A_{\sigma(s-1)}\cdots A_{\sigma(2)}A_{\sigma(1)}x_0 + \sum_{i=0}^{s} A_{\sigma(s)}A_{\sigma(s-1)}\cdots A_{\sigma(i+1)}B_{\sigma(i)}u(i)$$

Proof.-The case $s = 1$ is clear, we prove the result by induction. Assume the result for s; that is,

$$\Phi_\Gamma(x_0, \underline{\sigma}, \underline{u}) = A_{\sigma(s)}A_{\sigma(s-1)}\cdots A_{\sigma(2)}A_{\sigma(1)}x_0 + \sum_{i=0}^{s} A_{\sigma(s)}A_{\sigma(s-1)}\cdots A_{\sigma(i+1)}B_{\sigma(i)}u(i)$$

Consequently, by 2.5

$$\Phi_\Gamma(x_0, \underline{\sigma}\sigma(s+1), \underline{u}u(s+1)) = A_{\sigma(s+1)}A_{\sigma(s)}\cdots A_{\sigma(2)}A_{\sigma(1)}x_0 +$$
$$+A_{\sigma(s+1)}\left(\sum_{i=0}^{s} A_{\sigma(s)}A_{\sigma(s-1)}\cdots A_{\sigma(i+1)}B_{\sigma(i)}u(i)\right) + B_{\sigma(s+1)}u(s+1) =$$
$$= A_{\sigma(s+1)}A_{\sigma(s)}\cdots A_{\sigma(1)}A_{\sigma(0)}x_0 + \sum_{i=0}^{s+1} A_{\sigma(s)}A_{\sigma(s-1)}\cdots A_{\sigma(i+1)}B_{\sigma(i)}u(i)$$

and the result ■

Reachability is a central property of (dynamical) systems. In the switched linear case we research the set of internal states that can be reached by a given switched linear system Γ for any sequence of commands $\underline{\sigma}$ and any sequence of external inputs $\underline{u}$.

Definition 2.7 *Let $\Gamma : x(t+1) = A_{\sigma(t)}x(t) + B_{\sigma(t)}u(t)$ be a switched linear system. Let x_0 and ω be two internal states. We say that ω is switched reachable from initial state x_0 if there exists a chain of commands $\underline{\sigma}$ and a chain of inputs $\underline{u}$ such that*

$$\Phi_\Gamma(x_0, \underline{\sigma}, \underline{u}) = \omega$$

We will denote this fact by $x_0 \rightsquigarrow_{(\underline{\sigma},\underline{u})} \omega$ or simply by $x_0 \rightsquigarrow \omega$.

We say that Γ is reachable if for every pair of states x_1, x_2, one has that $x_1 \rightsquigarrow x_2$.

Interested reader is referred to [2], [10], [16] for the study of reachability and other related properties.

Next we introduce a way to study a given switched linear system Γ by using a new switched linear system $\widetilde{\Gamma}$ directly obtained from Γ. Main advantage is that $\widetilde{\Gamma} = (\widetilde{A}_\sigma, \widetilde{B})$ and all subsystems have same control matrix $\widetilde{B}$. This will yield a simplification of reachability calculations.

Let $\Gamma = (A_\sigma, B_\sigma)$ be a switched linear system, we define a new switched linear system $\widetilde{\Gamma} = \left(\widetilde{A}_\sigma, \widetilde{B}\right)$, where

$$\begin{aligned} \widetilde{A}_\sigma &= \begin{pmatrix} 0 & 0 \\ B_\sigma & A_\sigma \end{pmatrix} \\ \widetilde{B} &= \begin{pmatrix} Id & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Behavior of Γ and of $\widetilde{\Gamma}$ are closely related. In fact we have that reachability from zero-state is an equivalent notion in Γ and in $\widetilde{\Gamma}$. First we need to note an easy previous result.

Lemma 2.8

$$\Phi_{\widetilde{\Gamma}}\left(\begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix}, \sigma(0)\cdots\sigma(s), \begin{pmatrix} u(1) \\ 0 \end{pmatrix}\cdots\begin{pmatrix} u(s) \\ 0 \end{pmatrix}\begin{pmatrix} z \\ 0 \end{pmatrix}\right) =$$

$$= \begin{pmatrix} z \\ \Phi_\Gamma(\underline{0}, \sigma(1)\cdots\sigma(s), u(1)\cdots u(s)) \end{pmatrix}$$

Proof.- It is easily checked by induction on s ■

Theorem 2.9 *Switched Linear System Γ is reachable from $\underline{0}$ if and only if system $\widetilde{\Gamma}$ is reachable from $\begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix}$.*

Proof.- Suppose that system Γ is reachable from zero and let's prove that every internal state $\begin{pmatrix} \underline{\omega}_1 \\ \underline{\omega}_2 \end{pmatrix}$ of $\widetilde{\Gamma}$ can be reached from zero. Since Γ is reachable from zero it follows that

$$\underline{\omega}_2 = \Phi_\Gamma(\underline{0}, \sigma(1)\cdots\sigma(s), u(1)\cdots u(s))$$

Consequently

$$\Phi_{\widetilde{\Gamma}}\left(\begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix}, \sigma(0)\cdots\sigma(s), \begin{pmatrix} u(1) \\ 0 \end{pmatrix}\cdots\begin{pmatrix} u(s) \\ 0 \end{pmatrix}\begin{pmatrix} \omega_1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} \underline{\omega}_1 \\ \underline{\omega}_2 \end{pmatrix}$$

And we are done. The converse result is proved in a similar way ■

Definition 2.10 *Let $\Gamma : x(t+1) = A_{\sigma(t)}x(t) + Bu(t)$ be a switched linear system. Denote by $\mathrm{Reach}^s(\Gamma)$ the linear subspace of all reachable states from $\underline{0}$ with at most s commands; that is to say*

$$\mathrm{Reach}^s(\Gamma) = \{\underline{x} : \Phi_\Gamma(\underline{0}, \underline{\sigma}, \underline{u}) = \underline{x}; |\underline{\sigma}|, |\underline{u}| \le s\}$$

In the classical case of Example 2.2 it is well known that Γ is reachable if and only if $\mathrm{Reach}^n(\Gamma) = \mathbb{K}^n$. We will state the same result for the case of switched linear systems when $\mathbb{K}$ is an infinite field.

Obviously we have that $\mathrm{Reach}^s(\Gamma)$ is a subset $\mathrm{Reach}^{s+1}(\Gamma)$ for all s. But we can say anything more:

Lemma 2.11 $\mathrm{Reach}^s(\Gamma) = \mathrm{Reach}^{s+1}(\Gamma) \Rightarrow \mathrm{Reach}^{s+1}(\Gamma) = \mathrm{Reach}^{s+2}(\Gamma)$

Proof.- It is sufficient to prove that the following statement yields a contradiction:

$$\mathrm{Reach}^s(\Gamma) = \mathrm{Reach}^{s+1}(\Gamma) \subsetneq \mathrm{Reach}^{s+2}(\Gamma)$$

Let $\underline{x} \in \mathrm{Reach}^{s+2}(\Gamma) - \mathrm{Reach}^{s+1}(\Gamma)$ and assume that

$$\underline{x} = \Phi_\Gamma\left(\underline{0}, \sigma(0)\sigma(1)\cdots\sigma(s)\sigma(s+1), u(0)u(1)\cdots u(s)u(s+1)\right)$$

Then it follows that

$$\underline{x}' = \Phi_\Gamma\left(\underline{0}, \sigma(0)\sigma(1)\cdots\sigma(s), u(0)u(1)\cdots u(s)\right) \in \mathrm{Reach}^{s+1} = \mathrm{Reach}^s(\Gamma)$$

Consequently

$$\underline{x}' = \Phi_\Gamma\left(\underline{0}, \tau(0)\tau(1)\cdots\tau(s-1), v(0)v(1)\cdots v(s-1)\right)$$

for some $\underline{\tau}, \underline{v}$.

On the other hand $\underline{x} = \Phi_\Gamma\left(\underline{x}', \sigma(s+1), u(s+1)\right)$. Therefore

$$\underline{x} = \Phi_\Gamma\left(\underline{0}, \tau(0)\tau(1)\cdots\tau(s-1)\sigma(s+1), v(0)v(1)\cdots v(s-1)u(s+1)\right) \in \mathrm{Reach}^{s+1}(\Gamma)$$

which is a contradiction ■

First note that $\mathrm{Reach}^n(\Gamma)$ is not a linear subspace of the state space $\mathbb{K}^n$ but it is finite union of linear subspaces of $\mathbb{K}^n$ (see [13]). To be concise, if we denote by $\mathrm{Reach}_{\sigma(0)\cdots\sigma(s-1)}(\Gamma)$ the set of reachable states from zero by using the sequence of commands $\underline{\sigma}$ then

$$\mathrm{Reach}_{\underline{\sigma}}(\Gamma) = \mathrm{Im}(B, A_{\sigma(1)}B, ..., A_{\sigma(s-1)}\cdots A_{\sigma(1)}B)$$

and consequently

$$\mathrm{Reach}^n(\Gamma) = \bigcup_{|\underline{\sigma}|=n} \mathrm{Reach}_{\underline{\sigma}}(\Gamma)$$

Since we are working on infinite fields, a union of subspaces is the whole vector space if and only if one of involved subspaces is. Thus $\mathrm{Reach}^n(\Gamma) = \mathbb{K}^n$ if and only if $\mathrm{Reach}_{\underline{\sigma}}(\Gamma) = \mathbb{K}^n$ for some $\underline{\sigma}$.

On the other hand it is not difficult to check that $\mathrm{Reach}_{\underline{\sigma}}(\Gamma)$ is a linear subspace of $\mathrm{Reach}_{\underline{\sigma\tau}}(\Gamma)$ for all $\underline{\tau} \in \Sigma^*$. Therefore dimensions only can increase n times (all of them are subspaces of $\mathbb{K}^n$. Consequently the chain

$$\cdots \subseteq \mathrm{Reach}^s(\Gamma) \subseteq \mathrm{Reach}^{s+1}(\Gamma) \subseteq \cdots$$

stabilizes at index n. If a internal state cannot be reached using n commands then it can never be reached.

Above discussion is the proof of the following result:

Theorem 2.12 *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Let $\Gamma : x(t+1) = A_{\sigma(t)}x(t) + Bu(t)$ be a switched linear system. Then Γ is reachable from $\underline{0}$ if and only if* $\text{Reach}^n(\Gamma) = \mathbb{K}^n$

As main consequence we have the criterium of reachability of switched linear systems with common input matrix B in terms of reachability from zero.

Theorem 2.13 *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$. Let $\Gamma : x(t+1) = A_{\sigma(t)}x(t) + Bu(t)$ be a switched linear system. Then Γ is reachable if and only if Γ is reachable from $\underline{0}$*

Proof.- $\Rightarrow$ is straightforward. To prove the converse note that, by previous result we have that Γ is reachable from zero if and only if

$$\text{Reach}^n(\Gamma) = \bigcup_{|\underline{\sigma}|=n} \text{Reach}_{\underline{\sigma}}(\Gamma) = \mathbb{K}^n$$

A finite union of linear subespaces of $\mathbb{K}^n$ equals $\mathbb{K}^n$ if and only if one of them does. Hence we have that $\mathbb{K}^n = \text{Reach}_{\underline{\sigma}}(\Gamma)$ for some $\underline{\sigma}$ such that $|\underline{\sigma}| = n$. In particular,

$$x_2 - A_{\sigma(s)} \cdots A_{\sigma(0)} x_1 \in \text{Reach}_{\underline{\sigma}}(\Gamma)$$

Hence one has the equality

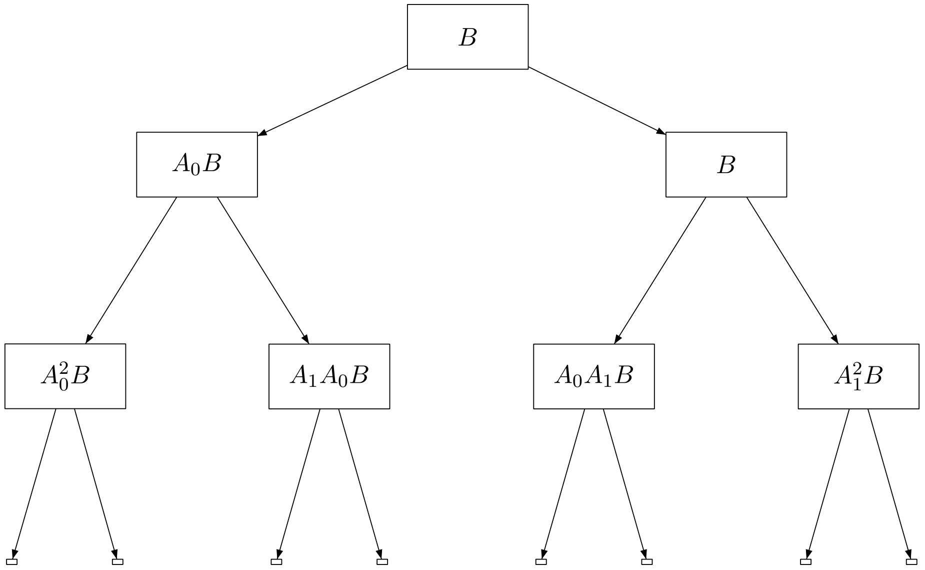
$$x_2 - A_{\sigma(s)} \cdots A_{\sigma(0)} x_1 = \Phi_\Gamma(\underline{0}, \underline{\sigma}, \underline{u})$$

which is equivalent to the equality

$$x_2 = \Phi_\Gamma(x_1, \underline{\sigma}, \underline{u})$$

Therefore $x_1 \rightsquigarrow x_2$ for all x_1, x_2 and Γ is reachable $\blacksquare$

Thus to obtain the reachable states of a switched linear systems it is sufficient to obtain $\sum_{k=0}^{n-1}(\#\Sigma)^k$ blocks that need to be adequately arranged. In the case of a switched linear system $\Gamma : x(t+1) = A_{\sigma(t)}x(t) + Bu(t)$ where $\sigma \in \{0, 1\}$ (i.e. two subsystems) we need to evaluate the following tree of block matrices:



We write down an explicit example for a switched linear system proposed in [16]:

Example 2.14 *Consider the three-dimensional single-input switched linear system ($\mathbb{K} = \mathbb{R}$) given by $\Gamma = (A_\sigma, B_\sigma, \Sigma = \{0,1\})$ where:*

$$A_0 = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, B_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 2 \\ 0 & 0 & -2 \end{pmatrix}, B_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Now, system $\widetilde{\Gamma}$ is given by

$$\widetilde{A_0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \widetilde{A_1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 \\ 1 & 0 & 0 & -2 \end{pmatrix}, \widetilde{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Now system is reachable if and only if it is reachable using a chain of at most 4 commands; Note the reverse indices from the sequence of commands and the indices of matrices. It is not difficult to complete the table and obtain that $\underline{\sigma} = 0010$ is a sequence of commands that reaches every state (using adequate input sequence $\underline{u}$) because

$$\operatorname{span}\{\widetilde{B}, \widetilde{A}_0\widetilde{B}, \widetilde{A}_1\widetilde{A}_0\widetilde{B}, \widetilde{A}_0\widetilde{A}_1\widetilde{A}_0\widetilde{B}\} = \mathbb{R}^4$$

.

3 Regular Switched Linear System

Some hybrid systems cannot be studied as switched systems (see above Example 2.4) because not all sequences of commands are allowed by the system. We introduce the notion of programmable switched linear system, which is closer to the notion of hybrid system (see [7])

Definition 3.1 *A programmable switched linear system is a pair* (Γ, L) *where* $\Gamma = (A_\sigma, B, \Sigma)$ *is a switched linear system and* $L \subseteq \Sigma^*$ *is a formal language of admissible chains (finite sequences) of commands of* Γ.

We say that (Γ, L) *is regular if* L *is a regular language.*

Note 3.2 *The classical case arise when* $\#\Sigma = 1$ *while usual switched linear systems are those where* $L = \Sigma^*$.

On the other hand note that there are some qualitative differences. Regular switched linear systems do generalize the notion of switched linear system; but on the other hand, key results Lemma 2.11 and Theorem 2.13 don't hold for regular switched linear.

Example 3.3 *We have seen for the Example 2.14 of previous section that system is reachable and that sequence of controls* 010 *gives* $\mathrm{Reach}_{010}(\Gamma) = \mathbb{R}^3$ *(in fact we have proven the equivalent equality* $\mathrm{Reach}_{0010}(\widetilde{\Gamma}) = \mathbb{R}^4$*). If we restrict our language of admissible chains of commands to, say*

$$L = 1^* = \{\varepsilon, 1, 11, 111, ...\}$$

then regular switched linear system (Γ, L) *is not reachable.*

Definition 3.4 *A programmable switched linear system* (Γ, L) *is reachable if for every internal states* $x_1, x_2 \in \mathbb{K}^n$ *one has* $x_1 \rightsquigarrow_{\underline{\sigma}} x_2$ *for some* $\underline{\sigma} \in L$.

We define the reachability language of a switched system Γ *as the set*

$$\pounds(\Gamma) = \{\underline{\sigma} \in \Sigma^* : \mathrm{Reach}_{\underline{\sigma}}(\Gamma) = \mathbb{K}^n\}$$

Note 3.5 *Note that if* $\underline{\sigma} \in \pounds(\Gamma)$ *then* $\underline{\sigma} \cdot \underline{\tau} \in \pounds(\Gamma)$ *for all* $\underline{\tau}$. *Hence the reachability language of a given programmable verifies the following formula for any formal language* $W \subseteq \Sigma^*$*:*

$$\pounds(\Gamma) \cdot W \subseteq \pounds(\Gamma)$$

Note 3.6 *A programmable switched linear system* $\Gamma = (A_\sigma, B, \Sigma)$ *is reachable if and only if* $\pounds(\Gamma) \neq \varnothing$

A possible algorithm to check reachability of a regular switched linear system would be as follows:

1. Consider matrix B.

2. Have spanned the whole vector space?

3. Check if the corresponding word is in L for all words spanning the whole vector space (if YES, then END).

4. Obtain next row of matrices.

5. GOTO (2)

Visiting every sequence is not an easy task in a general formal language L, even if L is also regular. On the other hand we do not have an upper bound to the size of the chains to be analyzed. Hence it is not assured the halting of the above procedure.

4 Local automata

Local languages are a class of regular languages (see definition bellow) having many computational advantages we are using in this section. Restricting our attention to local languages rather that regular languages is not too restrictive because the latter case can be reduced to the former case by linearization (see [6] for details). Thus in the sequel all regular languages we consider will be local languages.

Definition 4.1 *Let Σ be a finite alphabet. A formal language $L \subseteq \Sigma^*$ is local if and only if*

$$L - \{\varepsilon\} = (U\Sigma^* \cap \Sigma^* V) - (\Sigma^* W \Sigma^*)$$

where $U \subseteq \Sigma$ is a finite set of admissible-first-commands, $V \subseteq \Sigma$ is a finite set of admissible-last-commands and $W \subseteq \Sigma^2$ is a finite set of forbidden-pairs-of-commands.

Algorithm 4.2 *Let (Γ, L) be a local switched linear system where $\Gamma = (A_\sigma, B)$ and $L = (U\Sigma^* \cap \Sigma^* V) - (\Sigma^* W \Sigma^*)$ for some $U, V \subseteq \Sigma$ and some $W \subseteq \Sigma^2$. We describe an algorithm in order to study reachability of (Γ, L).*

Build the same tree than builded in section 2 by noticing that the root is B and the $(i+1)$th level is obtained from the ith level by expanding only through pairs of symbols not in W. That is to say, node $A_{\sigma(s)} \cdots A_{\sigma(1)} B$ expand to node $A_{\sigma(s+1)} A_{\sigma(s)} \cdots A_{\sigma(1)} B$ if and only if $\sigma(s)\sigma(s+1) \notin W$

After the completion of every level we check if $\sigma(s+1) \in V$ and if

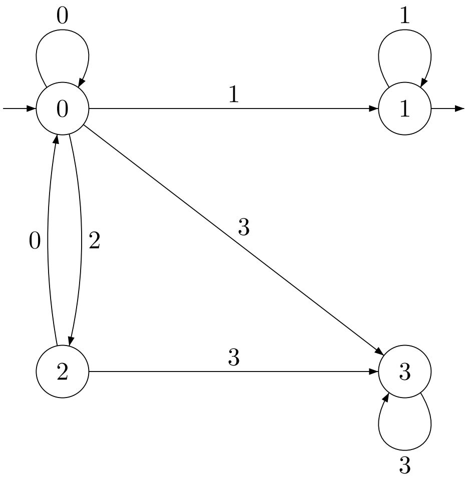
$$\mathrm{span}(B, A_{\sigma(1)}B, ..., A_{\sigma(s+1)} A_{\sigma(s)} \cdots A_{\sigma(1)} B) = \mathbb{K}^n$$

If it is true then system is reachable. Otherwise we expand the tree to the next level. Algorithm halts once level n is reached and every path to the root has been examined.

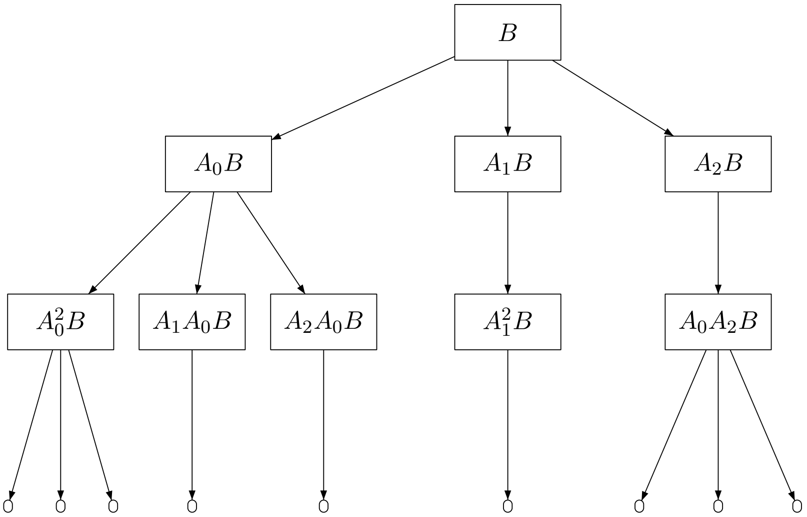
Example 4.3 *To conclude we expand the tree for a fixed local switched linear system. Let $\Sigma = \{0, 1, 2, 3\}$ and L be the local language defined by*

$$L = (U\Sigma^* \cap \Sigma^* V) - (\Sigma^* W \Sigma^*)$$

where $U = \{0\}, V = \{1\}, W = \{10, 12, 13, 21, 22, 30, 31, 32\}$. *That is to say, the local automaton associated to* L *is (see [6]):*



Hence for a regular local switched linear system where L *is the language of admissible sequences of commands we give the set of symbols where there doesn't exist any sequence of allowed commands to the final command* 1*: This set is* $\{3\}$. *Then we give the tree without using neither* 3 *nor any forbidden pair of symbols in* $W = \{10, 12, 13, 21, 22, 30, 31, 32\}$



Note that, in the study with local languages, dimensionality argument of Theorem 2.12 does apply. Hence we only need to expand the tree to level $n = \dim(\text{States})$ or to level where appear final symbols.

References

[1] M-P. Béal, J. Senellart, *On the bound of the synchronization delay of a local automaton*, Theoretical Computer Science **205** (1998) 297–306.

[2] D. Cheng, Y. Lin, Y. Wang, *Accesibility of switched linear systems*, IEEE Transactions on Automatic Control **51(9)** (2006) 1486–1491.

[3] Y. Fang, K. A. Loparo, *Stabilization of continuous-time jump linear systems*, IEEE Transactions on Automatic Control **47(10)** (2002) 1590–1603.

[4] A. Gabrielian, *The Theory of Interacting Local Automata*, Information and Control, **16** (1970) 360–377.

[5] M. Kim, R. McNaughton, R. McCloskey, *A Polynomial Time Algorithm for the Local Testability Problem od Deterministic Finite Automata*, IEEE Transactions on Computers, **40**(10) (1991) 1087–1093.

[6] M. V. Lawson, *Finite Automata*, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[7] P. J. Mosterman, *Hybrid Dynamic Systems: Modeling and Execution*, in: P. A. Fishwick (Ed.) Handbook of Dynamic System Modeling, Chapman & Hall/CRC , Boca Raton, FL, 2007.

[8] E. D. Sontag, *Interconnected Automata and Linear Systems: A Theoretical Framework in Discrete-Time*, in: R. Alur, T. A. Henzinger, E. D. Sontag (Eds.) Hybrid Systems III: Verification and Control, Springer , New York, 2007, 436–448.

[9] Z. Sun, *Sampling and control of switched linear systems*, Journal of The Franklin Institute, **31** (2004) 657–674.

[10] Z. Sun, *Switched Stability of Discrete-time Linear Hybrid Systems*, Proceedings 2007 IEEE International Conference on Control and Automation, (2007) 2777–2779.

[11] Z. Sun, S. S. Ge, *Analysis and synthesis of switched linear control systems*, Automatica **41** (2005) 181–195.

[12] Z. Sun, S. S. Ge, T. H. Lee, *Controllability and reachability criteria for switched linear systems*, Automatica, **38** (2002) 775–786.

[13] Z. Sun, D. Zheng, *On Reachability and Stabilization of Switched Linear Systems*, IEEE Transactions on Automatic Control, **46**(2) (2001) 291–295.

[14] R. Vidal, S. Soatto, Y. Ma, S. Sastry, *An Algebraic Geometric Approach to the Identification of a Class of Linear Hybrid Systems*, Proceedings 42nd. IEEE International Conference on Decision and Control, (2003) 167–172.

[15] G. Xie, L. Wang, *Reachability realization and stabilizability of switched linear discrete-time systems*, Journal of Mathematical Analysis and Applications, **280** (2003) 209–220.

[16] G. Xie, L. Wang, *Controllability and stabilizability of switched linear-systems*, Systems & Control Letters, **48** (2003) 135–155.

[17] G. Xie, D. Zheng, L. Wang, *Controllability of switched linear systems*, IEEE Transactions on Automatic Control, **47**(8) (2002) 1401–1405.

[18] Z. Yang, *An algebraic approach towards the controllability of controlled switching linear hybrid systems*, Automatica **38** (2002) 1221–1228.

*Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.*

New exact solutions of the variant of Boussinesq equations

Huaitang Chen[1,2,3]

[1] *Department of Mathematics, Linyi Normal University,Linyi 276005 P.R.China*

[2] *Department of Mathematics and IMS, Nanjing University, Nanjing, 210093 P.R.China*

[3] *Department of Applied Mathematics, Dalian University of Technology, Dalian 116024 P.R.China*

emails: `chenhuaitang@163.com`

Abstract

An elliptic equation method is presented for constructing exact travelling wave solutions of nonlinear partial differential equations(PDEs) in a unified way. With the aid of Maple, more new exact solutions are obtained for the variant of Boussinesq equations. This method can be applied to other PDEs.

Key words: Jacobian elliptic function, periodic solution, soliton solution
MSC 2000: AMS codes (35Q35)

1 Introduction

In recent years, directly searching for exact solutions of nonlinear PDEs has become more and more attractive partly due to the availability of computer symbolic systems like Maple and Mathematica, which allow us to perform some complicated and tedious algebraic calculations on computers as well as help us to find new exact solutions of PDEs. A number of methods have been presented, such as inverse scattering theory[1], Hirota's bilinear method[2], the truncated Painlevé expansion[3], homogeneous balance method[4], the hyperbolic tangent function series method[5], the sine-cosine method[6], and the Jacobi elliptic function method[7-8]. The purpose of this paper is to present an elliptic equation method and to solve the variant of Boussinesq equations as an example.

2 Our method based on the elliptic equation

The main idea of our method is to take full advantage of the elliptic equation that Jacobian elliptic functions satisfy. The desired elliptic equation reads

$$F'(\xi)^2 = AF(\xi) + BF(\xi)^2 + CF(\xi)^3, \tag{1}$$

where $F'(\xi) = dF(\xi)/d\xi$, and A, B, C are constants.

Case 1. If $\begin{cases} A = C = m^2 - 1 \\ B = 2(1 + m^2) \end{cases}$, then (1) has solution $F(\xi) = \frac{\mathrm{dn}^2\xi}{(1 \pm m\mathrm{sn}\xi)^2}$.

Case 2. If $\begin{cases} A = C = 1 - m^2 \\ B = 2(m^2 + 1) \end{cases}$, then (1) has solution $F(\xi) = \frac{\mathrm{cn}^2\xi}{(1 \pm \mathrm{sn}\xi)^2}$.

Case 3. If $\begin{cases} A = C = 1 \\ B = 2(1 - 2m^2) \end{cases}$, then (1) has solution $F(\xi) = \frac{\mathrm{sn}^2\xi}{(1 \pm \mathrm{cn}\xi)^2}$.

Consider a given PDE, say in two independent variables

$$H(u, u_t, u_x, u_{xx}, \cdots) = 0. \tag{2}$$

We assume that the solutions of Eq.(2) can be expressed in the form

$$u(x,t) = u(\xi) = a_0 + \sum_{i=1}^{n}[a_iF^i(\xi) + b_iF^{-i}(\xi)], \tag{3}$$

where, $\xi = kx + \omega t$, and n is a positive integer that can be determined by balancing the linear term of highest order with the nonlinear term in Eq.(2), and $k, \omega, a_0, a_i, b_i(i = 1, 2 \cdots, n)$ are parameters to be determined. Substituting (3) and (1) into Eq.(2) yields a set of algebraic equations for $a_0, a_i, b_i, (i = 1, 2, \cdots, n)$ and k, ω because all coefficients of F^i have to vanish. From these relations, $a_0, a_i, b_i, (i = 1, 2, \cdots, n)$ and k, ω can be determined.

In the following we illustrate our method by considering the variant of Boussinesq equations.

3 The variant of Boussinesq equations

As an example, we consider the variant of Boussinesq equations, which is written as:

$$u_t + guu_x + hv_x + pu_{xxt} = 0, \tag{4a}$$

$$v_t + ruv_x + svu_x + quxxx = 0, \tag{4b}$$

where, g, h, p, q and r are constants. The special cases of Eqs.(4a) and (4b) are studied by several authors.

In order to obtain travelling solutions of Eqs.(4a) and (4b), we Balance u_{xxx} with uu_x, and uv_x with u_{xxx} and get $n = 2$. Therefor, we choose the following ansatz:

$$u(x,t) = a_0 + a_1F(\xi) + a_2F(\xi)^2 + a_3/F(\xi) + a_4/F(\xi)^2, \tag{5a}$$

$$v(x,t) = b_0 + b_1F(\xi) + b_2F(\xi)^2 + b_3/F(\xi) + b_4/F(\xi)^2, \tag{5b}$$

where $\xi = kx + \omega t$.

Substituting (5a) and (5b) into (4a) and (4b) along with Eq.(1) and using Mathematica or Maple yields a system of equations w. r. t F^i. Setting the coefficients of F^i in the obtained system of equations to zero, we can deduce a set of equations with the respect unknowns $k, \omega, a_0, b_0, a_i, b_i (i = 1, 2, 3, 4)$. Solving these equations, we find

$a_2 = a_4 = b_2 = b_4 = 0$, k, ω are arbitrary nonzero constants.

$$a_0 = -\frac{p\omega^2 s + p\omega^2 r + p^2\omega^2k^2Bs + p^2\omega^2k^2Br + hqk^2g}{kp\omega g(s+r)}, (g \neq 0, p \neq 0, s + r \neq 0)$$

$$b_0 = -\frac{q(-rp\omega^2 s - r^2p\omega^2 + rp^2\omega^2k^2Bs - rqhk^2g + \omega^2gps + \omega^2gpr + k^2p^2\omega^2Bs^2)}{sp^2\omega^2(s+r)^2},$$

1.
$$a_1 = -\frac{3p\omega kC}{g}, b_1 = -\frac{3qk^2C}{s+r}, a_3 = b_3 = 0. \tag{6}$$

2.
$$a_3 = -\frac{3p\omega kA}{g}, b_3 = -\frac{3qk^2A}{s+r}, a_1 = b_1 = 0. \tag{7}$$

3.
$$a_1 = -\frac{3p\omega kC}{g}, b_1 = -\frac{3qk^2C}{s+r}, a_3 = -\frac{3p\omega kA}{g}, b_3 = -\frac{3qk^2A}{s+r}. \tag{8}$$

Substituting (6)-(8) into (5a) and (5b) and using the special solutions of Eq.(1), we obtain the Jacobian elliptic function solutions of Eqs. (4a) and (4b). For example, by using the solutions of Eq.(1) in Case 6, we obtain the following double periodic solutions of Eqs. (4a) and (4b).

$$u_1 = -\frac{p\omega^2 s + p\omega^2 r + hqk^2g + 2(1+m^2)(p^2\omega^2k^2s + p^2\omega^2k^2r)}{kp\omega g(s+r)} - \frac{3p\omega k(m^2-1)\mathrm{dn}^2\xi}{g(1 \pm m\mathrm{sn}\xi)^2}, \tag{9a}$$

$$v_1 = -\frac{q[-rp\omega^2 s - r^2p\omega^2 - rqhk^2g + \omega^2gps + \omega^2gpr + 2(1+m^2)(k^2p^2\omega^2s^2 + rp^2\omega^2k^2s)]}{sp^2\omega^2(s+r)^2}$$

$$-\frac{3qk^2(m^2-1)\mathrm{dn}^2\xi}{(s+r)(1 \pm m\mathrm{sn}\xi)^2}. \tag{9b}$$

$$u_2 = -\frac{p\omega^2 s + p\omega^2 r + hqk^2g + 2(1+m^2)(p^2\omega^2k^2s + p^2\omega^2k^2r)}{kp\omega g(s+r)} - \frac{3p\omega k(m^2-1)(1 \pm m\mathrm{sn}\xi)^2}{g\mathrm{dn}^2\xi}, \tag{10a}$$

$$v_2 = -\frac{q[-rp\omega^2 s - r^2p\omega^2 - rqhk^2g + \omega^2gps + \omega^2gpr + 2(1+m^2)(k^2p^2\omega^2s^2 + rp^2\omega^2k^2s)]}{sp^2\omega^2(s+r)^2}$$

$$-\frac{3qk^2(m^2-1)(1\pm m\mathrm{sn}\xi)^2}{(s+r)\mathrm{dn}^2\xi}. \tag{10b}$$

$$u_3=-\frac{p\omega^2s+p\omega^2r+hqk^2g+2(1+m^2)(p^2\omega^2k^2s+p^2\omega^2k^2r)}{kp\omega g(s+r)}$$

$$-\frac{3p\omega k(m^2-1)[(1\pm m\mathrm{sn}\xi)^4+\mathrm{dn}^4\xi]}{g\mathrm{dn}^2\xi(1\pm m\mathrm{sn}\xi)^2}, \tag{11a}$$

$$v_3=-\frac{q[-rp\omega^2s-r^2p\omega^2-rqhk^2g+\omega^2gps+\omega^2gpr+2(1+m^2)(k^2p^2\omega^2s^2+rp^2\omega^2k^2s)]}{sp^2\omega^2(s+r)^2}$$

$$-\frac{3qk^2(m^2-1)[(1\pm m\mathrm{sn}\xi)^4+\mathrm{dn}^4\xi]}{(s+r)\mathrm{dn}^2\xi(1\pm m\mathrm{sn}\xi)^2}. \tag{11b}$$

Here,$\xi=kx+\omega t$. By using the solutions of Eq.(1) in other cases, we can obtain other Jacobian elliptic function solutions of Eqs.(4a) and (4b). We omit these solutions from verbosity.

If $m\to 1$, then $\mathrm{sn}\xi\to\tanh\xi,\mathrm{cn}\xi\to\mathrm{sech}\xi,\mathrm{dn}\xi\to\mathrm{sech}\xi$, then these solutions can be degenerated as the soliton solutions. If $m\to 0$, then $\mathrm{sn}\xi\to\sin\xi,\mathrm{cn}\xi\to\cos\xi,\mathrm{dn}\xi\to 1$, then these solutions can be degenerated as the triangular solutions.

4 Conclusion and discussion

We present an elliptic equation method and solve the variant of Boussinesq equations. We obtain more new Jacobian elliptic function solutions. Our method is also a computerizable method, which allows us to perform complicated and tedious algebraic calculation on a computer. It can be applied to many other nonlinear PDEs. We are also aware of the fact that not all fundamental equations can be treated with our method. We are investigating how our method is further improved to treat more complicated and other kinds of nonlinear PDEs.

Acknowledgements

This research was supported by the National Natural Science Foundations of China(10671086 and 10571082), the Natural Science Foundations of Shandong Province (Y2007A26 and Y2005A12) and the Scientific Research Fund from Shandong Provincial Education Department (J07WH04).

References

[1] C. S. Gardner, J. M. Greene, M. D. Kruskal and R. M. Miura, *Method for solving the KdV equation*, Phys. Rev. Lett. **19** (1967) 1095–1097.

[2] X.B. Hu and W.X. Ma, *Application of Hirota's bilinear formalism to the Toeplitz latticesome special soliton-like solutions*, Phys. Lett. A **293** (2002) 161–165.

[3] J. Weiss, M. Taborand G. Garnevale, *The painlevé property for partial differential equations*, J. Math. Phys. **24** (1983) 522–526.

[4] M.L.Wang,Y.B.Zhou and Z.B.Li, *Applications of a homogeneous balance method to exact solutions of nonlinear equations in mathematical physics*, Phys.Lett.A **216** (1996) 67–75.

[5] B.Tibor,L.Béla,M.Csaba and U.Zsolt, *The hyperbolic tangent distribution family*, Powder Technology **97** (1998) 100–108.

[6] C.Yan, *A simple transformation for nonlinear waves*, Phys.Lett.A **224** (1996) 77–84.

[7] Z.T.Fu,S.K.Liu,S.D.Liuand Q.Zhao, *New Jacobi elliptic function expansion and new periodic solutions of nonlinear wave equations*, Phys.Lett.A **290** (2001) 72–76.

[8] H. T. Chen and H. Q. Zhang, *Improved Jacobin elliptic function method and its applications*, Chaos, Solitons and Fractals **15** (2003) 585–591.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

An almost second order uniform convergent computational technique for global solution and global normalized flux of singularly perturbed reaction-diffusion problems

C. Clavero[1], R. Bawa[2] and S. Natesan[3]

[1] *Department of Applied Mathematics, University of Zaragoza*

[2] *Department of Computer Science, Punjabi University*

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

emails: clavero@unizar.es, *rajesh_k_bawa@yahoo.com*, natesan@iitg.ernet.in

Abstract

In this paper, singularly perturbed reaction-diffusion two-point boundary value problems are considered. We are interested in constructing a computational technique, which is uniformly convergent for the global solution and the global normalized flux obtained from a classical cubic spline defined from the numerical solution at the mesh points. We have used the hybrid scheme, which is the combination of the cubic spline and the classical finite difference scheme developed by Natesan-Bawa-Clavero [5, 6], having almost second order uniform convergence at the nodal points when it is constructed on an appropriate piecewise–uniform Shishkin mesh. Using this scheme, we have defined the solution and the normalized flux on the whole domain. In the paper, we prove that the order of convergence of the global solution is same order of uniform convergence as that of above hybrid scheme. In addition, the global normalized flux is also almost second-order uniformly convergent in the whole domain. We have provided theoretical error bounds along with some numerical examples to show the efficiency of proposed technique for obtaining global solution and normalized flux.

Key words: singular perturbation problems, cubic spline, reaction-diffusion problem, Shishkin mesh, global solution, global normalized flux
MSC 2000: 65L10, 65L12, 65L20

1 Introduction

In this work we consider the following linear singularly perturbed reaction–diffusion two-point boundary-value problem (BVP)

$$Lu(x) \equiv -\varepsilon u''(x) + b(x)u(x) = f(x), \quad x \in D = (0,1), \tag{1}$$

$$u(0) = A, \quad u(1) = B, \tag{2}$$

where $\varepsilon > 0$ is a small parameter and b, f are sufficiently smooth functions such that $b^* \geq b(x) \geq \beta > 0$ on $\overline{D} = [0,1]$. Under these assumptions it is well known that the BVP (1-2) has a unique solution $u(x) \in \mathcal{C}^2(D) \cap \mathcal{C}(\overline{D})$, which in general has boundary layers of width $O(\sqrt{\varepsilon})$ at both end points (see [2, 8]).

In practice it is interesting to dispose of higher order uniform convergent methods for BVPs of the form (1-2), giving good approximations for any value of the diffusion parameter ε with a low computational cost. Some numerical methods having this property for singularly perturbed problems appear in [1, 3, 7], showing uniform convergence at the nodal points and also for the global solution. In [6], we have proposed an uniformly convergent scheme to solve the BVP (1-2) and obtain an almost second-order approximation for the global solution only in the boundary layer region, and for the normalized flux at the nodal points. Also, only computational results were given to approximate the global normalized flux.

In this paper, we extend the results of [6] by modifying the original Shishkin mesh in such a way that the global solution and the global normalized flux can be obtained in the whole domain, supported with complete theoretical proofs of their uniform convergence.

The paper is organized as follows: In Section 2, we define the numerical cubic spline constructed from the numerical solution at the mesh points obtained by the numerical scheme and we prove the uniform convergence for both the global solution and the global normalized flux. In Section 3 we display some numerical results obtained with the numerical method; these results corroborate in practice the theoretical results, showing the uniform convergence of the method and the order of convergence previously deduced.

Henceforth, C denotes any positive constant independent of the diffusion parameter ε and the discretization parameter N, which can take different values at different places.

2 The finite difference scheme: uniform convergence

Before developing the numerical method, we recall some standard results giving the asymptotic behaviour of the exact solution of a more general problem including the reaction-diffusion BVP (1-2). Appropriate bounds for the exact solution and also for its regular and singular components are showed (details of the proofs can be found in [4]). Let $y(x)$ be the solution of the following boundary value problem:

$$\begin{cases} -\varepsilon y''(x) + b(x)y(x) = g(x,\varepsilon), \quad x \in D, \\ y(0) = A, \quad y(1) = B, \end{cases}$$

where $b(x) \geq \beta > 0$, on $\overline{D}$, and the right hand side function $g(x,\varepsilon)$ satisfies the following bounds:

$$|g^{(k)}(x,\varepsilon)| \leq C\left(1 + \varepsilon^{-k/2} e(x,x,\beta,\varepsilon)\right), \; 0 \leq k \leq j,$$

where

$$e(\xi_1,\xi_2,\beta,\varepsilon) = \exp(-\sqrt{\beta}\xi_1/\sqrt{\varepsilon}) + \exp(-\sqrt{\beta}(1-\xi_2)/\sqrt{\varepsilon}), \tag{3}$$

and the value of j depends on the smoothness of data b and g. Then, the derivatives of y satisfy the following bound:

$$|y^{(k)}(x)| \le C\left(1+\varepsilon^{-k/2}e(x,x,\beta,\varepsilon)\right), \quad 0\le k\le j+1. \tag{4}$$

Further, we can decompose the solution $u(x)$ of (1-2) as $u(x)=v(x)+w(x)$, where v and w are the solutions of suitable problems (see [4] for further details), and they satisfy the following bounds:

$$|v^{(k)}(x)|\le C, \quad |w^{(k)}(x)|\le C\varepsilon^{-k/2}e(x,x,\beta,\varepsilon), \quad 0\le k\le j+1. \tag{5}$$

The functions v and w are respectively known as the regular and the singular components of the exact solution u.

To construct the finite difference scheme, first we define the original piecewise uniform Shsihkin mesh as follows. The domain $\overline{D}$ is divided into three subintervals as $\overline{D}=[0,\sigma)\cup[\sigma,1-\sigma]\cup(1-\sigma,1]$, for some σ such that $0<\sigma\le 1/4$. On the subintervals $[0,\sigma]$ and $[1-\sigma,1]$ a uniform mesh with $N/4$ mesh intervals are placed, while $[\sigma,1-\sigma]$ has a uniform mesh with $N/2$ mesh intervals. It is obvious that the mesh is uniform when $\sigma=1/4$ and it is fitted to the problem by choosing σ as the following function of N, ε and σ_0

$$\sigma=\min\left\{1/4,\sigma_0\sqrt{\varepsilon}\ln N\right\}, \tag{6}$$

where σ_0 is a constant to be fixed later. The mesh size in the region $[\sigma,1-\sigma]$ is $H=2(1-2\sigma)/N$, and in the regions $[0,\sigma]$, $[1-\sigma,1]$ it is $h=4\sigma/N$. Below we denote $h_{i+1}=x_{i+1}-x_i,\ i=0,1,\cdots,N-1$. On this mesh we consider the finite difference scheme (see [6])

$$L^N U_i^N \equiv r_i^- U_{i-1}^N + r_i^c U_i^N + r_i^+ U_{i+1}^N = q_i^- f_{i-1} + q_i^c f_i + q_i^+ f_{i+1}, 1\le i\le N-1 \tag{7}$$

along with the boundary conditions $U_0^N=A$ and $U_N^N=B$, where, for indices $i=1,\cdots,N/4-1$ and also $3N/4+1,\cdots,N-1$, the coefficients are given by

$$\begin{cases} r_i^- = \dfrac{-3\varepsilon}{h_i(h_i+h_{i+1})}+\dfrac{h_i}{2(h_i+h_{i+1})}b_{i-1}, \quad r_i^c=\dfrac{3\varepsilon}{h_ih_{i+1}}+b_i,\\ r_i^+ = \dfrac{-3\varepsilon}{h_{i+1}(h_i+h_{i+1})}+\dfrac{h_{i+1}}{2(h_i+h_{i+1})}b_{i+1},\\ q_i^- = \dfrac{h_i}{2(h_i+h_{i+1})}, \quad q_i^c=1, \quad q_i^+=\dfrac{h_{i+1}}{2(h_i+h_{i+1})}, \end{cases} \tag{8}$$

and for $i=N/4,\cdots,3N/4$, the coefficients are given by

$$\begin{cases} r_i^- = \dfrac{-2\varepsilon}{h_i(h_i+h_{i+1})}, \quad r_i^c=\dfrac{2\varepsilon}{h_ih_{i+1}}+b_i, \quad r_i^+=\dfrac{-2\varepsilon}{h_{i+1}(h_i+h_{i+1})},\\ q_i^-=0, \quad q_i^c=1, \quad q_i^+=0. \end{cases} \tag{9}$$

In [5], we have obtained the following ε–uniform error estimate for the above difference scheme at the mesh points.

Theorem 1 *Let $u(x)$ be the solution of (1-2) and U^N be the numerical solution of the hybrid finite difference scheme (7)-(9). Then, we have the following ε–uniform error bound:*

$$|u(x_i) - U_i^N| \le C\left(N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right), \quad i = 0, 1, \cdots, N, \tag{10}$$

One can easily observe that if $\sigma_0 \ge 2/\sqrt{\beta}$, then the numerical scheme (7)-(9) is uniformly convergent of second-order up to a logarithmic factor.

Let $\overline{D}^N \equiv \{x_i : 0 = x_0 < \cdots < x_N = 1\}$ be the mesh. Then, for given values $u(x_0), u(x_1), \cdots, u(x_N)$ of a function $u(x)$ at the nodal points $x_0, x_1, \ldots, x_N$, it is well known that there exists an interpolating cubic spline $s(x)$ given by

$$s(x) = \frac{(x_{i+1}-x)^3}{6h_{i+1}}M_i + \frac{(x-x_i)^3}{6h_{i+1}}M_{i+1} + \left(u_i - \frac{h_{i+1}^2}{6}M_i\right)\left(\frac{x_{i+1}-x}{h_{i+1}}\right) + $$
$$+\left(u_{i+1} - \frac{h_{i+1}^2}{6}M_{i+1}\right)\left(\frac{x-x_i}{h_{i+1}}\right), \; x_i \le x \le x_{i+1}, \; i = 0, \cdots, N-1, \tag{11}$$

where $u_i = u(x_i), M_i = u''(x_i), i = 0, \cdots, N$. Using this cubic spline an approximation to the the global normalized flux can be defined as

$$\sqrt{\varepsilon}s'(x) = -\sqrt{\varepsilon}\; M_i\frac{(x_{i+1}-x)^2}{2h_{i+1}} + \sqrt{\varepsilon}\; M_{i+1}\frac{(x-x_i)^2}{2h_{i+1}} + \sqrt{\varepsilon}\frac{u_{i+1}-u_i}{h_{i+1}} -$$
$$-\sqrt{\varepsilon}\frac{(M_{i+1}-M_i)}{6}h_{i+1}, \quad x_i \le x \le x_{i+1}, \; i = 0, \cdots, N-1.$$

To calculate the numerical cubic spline we use the discrete solution $U_0, U_1, \cdots U_N$ at mesh points. Defining $\overline{M}_i = (b_iU_i - f_i)/\varepsilon$, $i = 0, \cdots N$, we construct the cubic spline as

$$S(x) = \frac{(x_{i+1}-x)^3}{6h_{i+1}}\overline{M}_i + \frac{(x-x_i)^3}{6h_{i+1}}\overline{M}_{i+1} + \left(U_i - \frac{h_{i+1}^2}{6}\overline{M}_i\right)\left(\frac{x_{i+1}-x}{h_{i+1}}\right) + $$
$$+\left(U_{i+1} - \frac{h_{i+1}^2}{6}\overline{M}_{i+1}\right)\left(\frac{x-x_i}{h_{i+1}}\right), \quad x_i \le x \le x_{i+1}, i = 0, \cdots, N-1. \tag{12}$$

Note that using this spline we can obtain a numerical approximation to the exact solution at any point in $[0, 1]$. Also, we can obtain an approximation to the normalized flux as follows:

$$\sqrt{\varepsilon}S'(x) = -\sqrt{\varepsilon}\; \overline{M}_i\frac{(x_{i+1}-x)^2}{2h_{i+1}} + \sqrt{\varepsilon}\; \overline{M}_{i+1}\frac{(x-x_i)^2}{2h_{i+1}} + \sqrt{\varepsilon}\frac{U_{i+1}-U_i}{h_{i+1}} -$$
$$-\sqrt{\varepsilon}\frac{(\overline{M}_{i+1}-\overline{M}_i)}{6}h_{i+1}, \quad x_i \le x \le x_{i+1}, \; i = 0, \cdots, N-1.$$

Lemma 2 *Let $u(x)$ be the solution of (1)-(2) and $s(x)$ be the spline given by (11). Then, for $x \in [x_i, x_{i+1}]$, $i = 0, 1, \cdots, N-1$, we have the following estimate:*

$$|s(x) - u(x)| \le Ch_{i+1}^3\left(1 + \varepsilon^{-3/2}e(x_i, x_i, \beta, \varepsilon)\right). \tag{13}$$

Proof. Let $x = x_i + \theta h_{i+1}$ with $0 \leq \theta \leq 1$. Using that $u''_j = M_j = (b_j u_j - f_j)/\varepsilon$, $j = i, i+1$, from (11), the Taylor expansions and the bounds (4) for the derivatives of the exact solution, it is straightforward to obtain

$$|s(x) - u(x)| \leq Ch_{i+1}^3 |u^3(\xi)| \leq Ch_{i+1}^3 \left(1 + \varepsilon^{-3/2} e(x_i, x_i, \beta, \varepsilon)\right). \qquad \blacksquare$$

Lemma 3 *Let $s(x)$ be the spline given in (11) and $S(x)$ be the numerical spline given in (12). Then, for $x \in [x_i, x_{i+1}]$, $i = 0, 1, \cdots, N-1$, the following result holds*

$$|s(x) - S(x)| \leq C\left(1 + b^* \frac{h_{i+1}^2}{\varepsilon}\right) \max\{|u_i - U_i|, |u_{i+1} - U_{i+1}|\}. \tag{14}$$

Proof. Let $x = x_i + \theta h_{i+1}$, where $0 \leq \theta \leq 1$. From expressions (11) and (12) for exact and numerical splines, using that $M_j = (b_j u_j - f_j)/\varepsilon$, $j = i, i+1$, $\overline{M}_j = (b_j U_j - f_j)/\varepsilon$, $j = i, i+1$, and taking the absolute values, one can easily obtain the required result. $\blacksquare$

Using these two lemmas we are in a position to obtain the uniform convergence of the global solution based on the numerical cubic spline.

Theorem 4 *Let $u(x)$ be the solution of (1)-(2) and $S(x)$ be the spline given in (12). Then, we have the following error bound:*

$$|S(x) - u(x)| \leq \left(N^{-2} \ln^2 N + N^{3-\sqrt{\beta}\sigma_0} \ln^3 N\right). \tag{15}$$

Proof. First we assume that the mesh is uniform, *i.e.*, $\sigma = 1/4$ and $1/\sqrt{\varepsilon} \leq C \ln N$. Then, from (13), (14) and Theorem 1 it is straightforward to obtain that

$$|S(x) - u(x)| \leq C\left(N^{-3} \ln^3 N + N^{-\sqrt{\beta}\sigma_0}\right). \tag{16}$$

In the second place, we assume that $1/4 > \sigma_0 \sqrt{\varepsilon} \ln N$, which is the most interesting case in practice. We only give the details for $x \leq 1/2$; from the symmetry of the boundary layers at both end points, the proof for $x \geq 1/2$ is the same. We distinguish several cases depending on the location of the mesh point x_i. First, when $i = 0, 1, \cdots, N/4 - 1$, the mesh point x_i is in the boundary layer region. Now $h_{i+1} = 4N^{-1}\sqrt{\varepsilon}\sigma_0 \ln N$; from (13) it follows that

$$|s(x) - u(x)| \leq CN^{-3} \ln^3 N,$$

and from (14) and Theorem 1 we have

$$|s(x) - S(x)| \leq C\left(N^{-2} \ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right).$$

From these two previous bounds we obtain

$$|S(x) - u(x)| \leq C\left(N^{-2} \ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right). \tag{17}$$

The second case is for $i = N/4 + 1, \cdots, N/2 - 1$, *i.e.*, the mesh point x_i is outside the boundary layer region. To find appropriate bounds of the error, we analyze two different subcases: $H \leq \sqrt{\varepsilon}$ and $H > \sqrt{\varepsilon}$. For the first one, from (13) we deduce

$$|s(x) - u(x)| \leq C\left(N^{-3} + e(x_i, x_i, \beta, \varepsilon)\right) \leq C\left(N^{-3} + N^{-\sqrt{\beta}\sigma_0}\right),$$

and from (14) and Theorem 1 we obtain

$$|s(x) - S(x)| \leq C\left(N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right).$$

From the two previous bounds it follows

$$|S(x) - u(x)| \leq C\left(N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right).$$

In the second subcase the proof is different; now, we do not use Lemma 3 and we follow similar ideas to these ones developed in [6]. Using the uniform stability of the hybrid scheme we have

$$|s(x) - S(x)| \leq C\left(1 + b^*\frac{h_{i+1}^2}{\varepsilon}\right)|\tau_i|, \tag{18}$$

where τ_i is the local error at x_i. In [5], it was proved that

$$\tau_i = \frac{\varepsilon}{H^2}\left(R_3(x_i, x_{i+1}, u) + R_3(x_i, x_{i-1}, u)\right), \tag{19}$$

where R_n denotes the remainder of Taylor expansion and therefore

$$|s(x) - S(x)| \leq C\left(R_3(x_i, x_{i+1}, u) + R_3(x_i, x_{i-1}, u)\right).$$

Using the integral form of the remainder for the Taylor expansion, integrating by parts and taking into account that $e(x_j, x_j, \beta, \varepsilon) \leq N^{-\sqrt{\beta}\sigma_0}$, $j = i-1, i, i+1$, we can obtain

$$|s(x) - S(x)| \leq C\left(N^{-4} + N^{-\sqrt{\beta}\sigma_0}\right).$$

On the other hand, using that $x_i \geq \sigma + H$ and the definition of the transition parameter σ, we have

$$\left(H\varepsilon^{-1/2}\right)^l e(x_i, x_i, \beta, \varepsilon) \leq CN^{-\sqrt{\beta}\sigma_0},\ l = 0, \cdots, 4, \tag{20}$$

and therefore

$$|s(x) - u(x)| \leq C\left(N^{-3} + N^{-\sqrt{\beta}\sigma_0}\right).$$

From previous bounds we deduce

$$|S(x) - u(x)| \leq C\left(N^{-3} + N^{-\sqrt{\beta}\sigma_0}\right).$$

The last case corresponds to the mesh point $x_{N/4} = \sigma$. We only give the details for the more difficult case $H > \sqrt{\varepsilon}$. Following the original idea of Surla (see [11]) for mesh modification, we define $\overline{H} = \sqrt{\varepsilon/\beta}N\ln N$. Then, if $H/2 \leq \overline{H}$ the Shishkin mesh remains unchanged. From (13) we can obtain

$$|s(x) - u(x)| \leq C\left(N^{-3} + N^{3-\sqrt{\beta}\sigma_0}\ln^3 N\right).$$

On the other hand, we use again (18), where the local truncation error can be bounded by

$$|\tau_{N/4}| \leq |r^+_{N/4}R_2(x_{N/4}, x_{N/4+1}, u)| + |r^-_{N/4}R_2(x_{N/4}, x_{N/4-1}, u)|.$$

Following the way of proof of [5], we can prove that

$$(H^2/\varepsilon)|r^+_{N/4}R_2(x_{N/4},x_{N/4+1},u)| \leq C\left(N^{-3}+N^{-\sqrt{\beta}\sigma_0}\right),$$

and

$$(H^2/\varepsilon)|r^-_{N/4}R_2(x_{N/4},x_{N/4-1},u)| \leq C\left(N^{-3}+N^{-\sqrt{\beta}\sigma_0}\ln^2 N\right).$$

From previous bounds we have

$$|s(x)-S(x)| \leq C\left(N^{-3}+N^{-\sqrt{\beta}\sigma_0}\ln^2 N\right),$$

and therefore

$$|S(x)-u(x)| \leq C\left(N^{-3}+N^{-\sqrt{\beta}\sigma_0}\ln^3 N\right).$$

Finally, when $H/2 > \overline{H}$ we modify the original Shishkin mesh adding the mesh points $\overline{x}_{N/4} = x_{N/4}+\overline{H}$ and $\overline{x}_{3N/4} = x_{3N/4}-\overline{H}$ and we put $N = N+2$ and enumerate the points. Using again the local error we can prove

$$|s(x)-S(x)| \leq C\left(N^{-3}+N^{-\sqrt{\beta}\sigma_0}\ln^2 N\right),$$

and considering $x \in [\sigma, \sigma+\overline{H}]$ we obtain

$$|s(x)-u(x)| \leq C\left(N^{-3}+N^{3-\sqrt{\beta}\sigma_0}\ln^3 N\right),$$

and therefore

$$|S(x)-u(x)| \leq C\left(N^{-3}+N^{3-\sqrt{\beta}\sigma_0}\ln^3 N\right). \qquad \blacksquare$$

Remark 5 *From Theorem 4 we see that by taking the constant σ_0 such that $\sqrt{\beta}\sigma_0 \geq 5$ the global solution has order of uniform convergence $O(N^{-2}\ln^3 N)$; in that sense the result is optimal because we have the same order of uniform convergence as at the mesh points.*

Now we prove the uniform convergence of the normalized flux.

Lemma 6 *Let $u(x)$ be the solution of (1-2) and $s(x)$ be the spline given by (11) and $S(x)$ be the spline given in (12). Then, for $x \in [x_i, x_{i+1}]$, $i = 0, 1, \cdots, N-1$, the following bounds hold:*

$$\sqrt{\varepsilon}|s'(x)-u'(x)| \;\leq\; C\sqrt{\varepsilon}h^3_{i+1}\left(1+\varepsilon^{-2}e(x_i,x_i,\beta,\varepsilon)\right). \tag{21}$$

$$\sqrt{\varepsilon}|s'(x)-S'(x)| \leq C\left(\frac{\sqrt{\varepsilon}}{h_{i+1}}+b^*\frac{h_{i+1}}{\sqrt{\varepsilon}}\right)\max\{|u_i-U_i|, |u_{i+1}-U_{i+1}|\}. \tag{22}$$

Proof. The proof is straightforward using (11), (12) and the Taylor expansions. ■

Theorem 7 *Let $\sqrt{\varepsilon}u'(x)$ be the normalized flux of (1-2) and $\sqrt{\varepsilon}S'(x)$ be the normalized flux obtained from cubic spline approximations. Then, we have*

$$\sqrt{\varepsilon}|S'(x)-u'(x)| \leq \begin{cases} C(N^{-2}\ln^3 N + N^{3-\sqrt{\beta}\sigma_0}\ln^3 N), \textit{ if } \ N^{-1} > \sqrt{\varepsilon} \\ (N^{-1}\sqrt{\varepsilon}\ln^2 N + N^{1-\sqrt{\beta}\sigma_0}), \textit{ if } \ N^{-1} \leq \sqrt{\varepsilon} \end{cases} \tag{23}$$

Proof. Let $x = x_i + \theta h_{i+1}, 0 \leq \theta \leq 1$. From (11) and (12), we have

$$S''(x) - u''(x) = (1-\theta)(\overline{M}_i - u''(x_i)) + \theta(\overline{M}_{i+1} - u''(x_{i+1})).$$

Using that $u''(x_j) = M_j = (b_j u_j - f_j)/\varepsilon, j = i, i+1$, $\overline{M}_j = (b_j U_j - f_j)/\varepsilon, j = i, i+1$ and (10) we obtain

$$|u''(x) - S''(x)| \leq C\varepsilon^{-1}\left(N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right) \tag{24}$$

Below we use in different places that for any function $g(x)$ it holds

$$\parallel g' \parallel_{[a,b]} \leq \frac{2 \parallel g \parallel_{[a,b]}}{(b-a)} + (b-a) \parallel g'' \parallel_{[a,b]} . \tag{25}$$

First we assume that the mesh is uniform, *i.e.*, $1/4 \leq \sigma_0\sqrt{\varepsilon}\ln N$. Taking $g(x) = u(x) - S(x)$, $[a,b] = [0,1]$ in (25), using (16) and (24), it follows

$$\sqrt{\varepsilon}|u'(x) - S'(x)| \leq C[\varepsilon^{1/2}(N^{-3}\ln^3 N + N^{-\sqrt{\beta}\sigma_0})+$$
$$+\varepsilon^{-1/2}(N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0})] \leq \left(CN^{-2}\ln^3 N + N^{-\sqrt{\beta}\sigma_0}\right).$$

In second case, it holds that $1/4 > \sigma_0\sqrt{\varepsilon}\ln N$. Again we only consider the details for $x \leq 1/2$. In the boundary layer region, $i = 0, 1, \cdots, N/4 - 1$, we have $h_{i+1} = 4N^{-1}\sqrt{\varepsilon}\sigma_0\ln N$ and $x \in [0,\sigma]$. From (25) and (20), with same g as before and now $[a,b] = [0,\sigma]$, we get

$$\sqrt{\varepsilon}|u'(x) - S'(x)| \leq C\left(N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0}\right).$$

Outside the boundary layer region, $i = N/4 + 1, \cdots, N/2 - 1$, we distinguish two subcases: $H \leq \sqrt{\varepsilon}$ and $H > \sqrt{\varepsilon}$. For the first one, from (21) and (22) and Theorem 1, we have

$$\sqrt{\varepsilon}|s'(x) - u'(x)| \leq C(\varepsilon^{1/2}N^{-3} + N^{-\sqrt{\beta}\sigma_0}),$$
$$\sqrt{\varepsilon}|s'(x) - S'(x)| \leq C\left[\varepsilon^{1/2}(N^{-1}\ln^2 N + N^{1-\sqrt{\beta}\sigma_0}) + (N^{-2}\ln^2 N + N^{-\sqrt{\beta}\sigma_0})\right].$$

From previous bounds it follows,

$$\sqrt{\varepsilon}|u'(x) - S'(x)| \leq C\varepsilon^{1/2}\left(N^{-1}\ln^2 N + N^{1-\sqrt{\beta}\sigma_0}\right).$$

For the second one, $H > \sqrt{\varepsilon}$, using (20) in (21), we obtain

$$\sqrt{\varepsilon}|s'(x) - u'(x)| \leq C(N^{-3} + N^{-\sqrt{\beta}\sigma_0}). \tag{26}$$

On the other hand, following [6], we can deduce that

$$\sqrt{\varepsilon}|s'(x) - S'(x)| \leq C(1 + b^* \frac{h_{i+1}}{\sqrt{\varepsilon}})|\tau_i|,$$

where τ_i is the local error at x_i given in (19). The same analysis that in Theorem 4 gives

$$\sqrt{\varepsilon}|s'(x) - S'(x)| \leq C\left(N^{-4} + N^{-\sqrt{\beta}\sigma_0}\right). \tag{27}$$

From (26) and (27) it follows

$$\sqrt{\varepsilon}|S'(x) - u'(x)| \leq C(N^{-3} + N^{-\sqrt{\beta}\sigma_0}).$$

The last case is when the mesh point is the transition point, *i.e.*, $x \in [x_{N/4}, x_{N/4+1}]$. As in Theorem 4 we only give the details for the case $H > \sqrt{\varepsilon}$. Defining $\overline{H} = \sqrt{\varepsilon/\beta} N \ln N$, and considering the same modified Shishkin mesh as before, for $H/2 \leq \overline{H}$, from (21) we have

$$\sqrt{\varepsilon}|s'(x)-u'(x)| \leq C\left(H^3\sqrt{\varepsilon} + \overline{H}^3 \varepsilon^{-3/2} e(x_i, x_i, \beta, \varepsilon)\right) \leq C\left(N^{-3}\sqrt{\varepsilon} + N^{3-\sqrt{\beta}\sigma_0} \ln^3 N\right).$$

On the other hand, using again the local error, we can obtain

$$\sqrt{\varepsilon}|s'(x) - S'(x)| \leq C\left(N^{-3} + N^{-\sqrt{\beta}\sigma_0} \ln^2 N\right).$$

and therefore

$$\sqrt{\varepsilon}|S'(x) - u'(x)| \leq C\left(N^{-3} + N^{3-\sqrt{\beta}\sigma_0} \ln^3 N\right),$$

Finally, when $H/2 > \overline{H}$ and $x \in [\sigma, \sigma + \overline{H}]$, using the same technique as before we can obtain

$$\sqrt{\varepsilon}|s'(x) - u'(x)| \leq C\left(N^{-3}\sqrt{\varepsilon} + N^{3-\sqrt{\beta}\sigma_0} \ln^3 N\right),$$

and

$$\sqrt{\varepsilon}|s'(x) - S'(x)| \leq C\left(N^{-3} + N^{-\sqrt{\beta}\sigma_0} \ln^2 N\right).$$

and therefore

$$\sqrt{\varepsilon}|S'(x) - u'(x)| \leq C\left(N^{-3} + N^{3-\sqrt{\beta}\sigma_0} \ln^3 N\right).$$

Clubbing all the cases the result follows. ■

Remark 8 *From Theorem 7, we see that by taking the constant σ_0 such that $\sqrt{\beta}\sigma_0 \geq 5$, the global normalized flux has almost second order of uniform convergence except for $N^{-1} \leq \sqrt{\varepsilon}$, which is not a practical case. So, practically, the obtained global normalized flux is almost second order convergent. Our computational results also confirms the same fact.*

3 Numerical Experiments

The example that we consider is

$$\begin{cases} -\varepsilon u''(x) + (e^x + \sin(x) - x - x^3)u(x) = \cos(x) + x^2 - e^x + 1, x \in (0,1), \\ u(0) = 0, \quad u(1) = 2. \end{cases} \tag{28}$$

We are only interested in the errors outside the mesh points. To obtain an approximation to the maximum errors and the rates of convergence, we use a variant of the double mesh principle (see [9, 10]). We calculate the numerical solution U^N on D^N and the numerical solution $\widetilde{U}^N$ on the mesh $\widetilde{D}^N$ where the transition parameter is now given by

$$\widetilde{\sigma} = \min\left\{1/4, \sigma_0\sqrt{\varepsilon}\ln(N/2)\right\}.$$

Then, the errors at midpoints $x = (x_i + x_{i+1})/2$, of the corresponding Shishkin mesh, are calculated by

$$E_\varepsilon^N = \max_x |S^N(x) - \widetilde{S}^{2N}(x)|, \quad E^N = \max_\varepsilon E_\varepsilon^N,$$

where S_N and $\widetilde{S}_{2N}$ are the splines defined by (12) on the meshes D^N and $\widetilde{D}^{2N}$ respectively. Using these errors, the numerical orders of convergence and the uniform orders of convergence are given by

$$p_\varepsilon^N = \log_2\left(E_\varepsilon^N(x)/E_\varepsilon^{2N}(x)\right), \quad p^N = \log_2\left(E^N/E^{2N}\right).$$

To permit that the maximum errors stabilize, we take, for the diffusion parameter, the set of values $\varepsilon = 2^0, 2^{-2}, 2^{-4}, \cdots, 2^{-48}$.

From table 1, it can be seen the almost second order of uniform convergence for the global solution, in agreement with Theorem 4.

Table 1: *Maximum errors for the solution at midpoints,on the modified Shishkin mesh, and rates of convergence taking $\sigma_0 = 5$ for problem (28)*

ε/N	16	32	64	128	256	512	1024	2048
2^0	5.3821E-5	1.2776E-5	3.0927E-6	7.5883E-7	1.8786E-7	4.6730E-8	1.1651E-8	2.9113E-9
	2.0747	2.0465	2.0270	2.0141	2.0073	2.0039	2.0007	
2^{-8}	4.6780E-2	9.9632E-3	2.3483E-3	5.8148E-4	1.4481E-4	3.6184E-5	9.0446E-6	2.2610E-6
	2.2312	2.0850	2.0138	2.0056	2.0007	2.0002	2.0001	
2^{-16}	1.2953E+0	3.9210E-1	9.9217E-2	2.3251E-2	7.4105E-3	2.2821E-3	6.9580E-4	2.1009E-4
	1.7240	1.9826	2.0933	1.6496	1.6992	1.7136	1.7276	
2^{-24}	1.2966E+0	3.9267E-1	9.9396E-2	2.3293E-2	7.4237E-3	2.2860E-3	6.9700E-4	2.1045E-4
	1.7233	1.9821	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-32}	1.2967E+0	3.9271E-1	9.9407E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9707E-4	2.1048E-4
	1.7233	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-40}	1.2967E+0	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	1.7233	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-48}	1.2967E+0	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	1.7233	1.9820	2.0933	1.6497	1.6993	1.7136	1.7276	
E^N	1.2967E+0	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
p^N	1.7233	1.9820	2.0933	1.6497	1.6993	1.7136	1.7276	

Table 2: *Maximum errors for the solution at midpoints,on the original Shishkin mesh, and rates of convergence taking $\sigma_0 = 5$ for problem (28)*

ε/N	16	32	64	128	256	512	1024	2048
2^0	5.3821E-5	1.2776E-5	3.0927E-6	7.5883E-7	1.8786E-7	4.6730E-8	1.1651E-8	2.9113E-9
	2.0747	2.0465	2.0270	2.0141	2.0073	2.0039	2.0007	
2^{-8}	4.6780E-2	9.9632E-3	2.3483E-3	5.8148E-4	1.4481E-4	3.6184E-5	9.0446E-6	2.2610E-6
	2.2312	2.0850	2.0138	2.0056	2.0007	2.0002	2.0001	
2^{-16}	1.2953E+0	3.9210E-1	9.9217E-2	2.3251E-2	7.4105E-3	2.2821E-3	6.9580E-4	2.1009E-4
	1.7240	1.9826	2.0933	1.6496	1.6992	1.7136	1.7276	
2^{-24}	1.2966E+0	3.9267E-1	9.9396E-2	2.3293E-2	7.4237E-3	2.2860E-3	6.9700E-4	2.1045E-4
	1.7233	1.9821	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-32}	2.0490E+0	3.9271E-1	9.9407E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9707E-4	2.1048E-4
	2.3834	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-40}	3.2817E+1	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	6.3848	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-42}	6.5636E+1	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	7.3849	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-44}	1.3127E+2	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	8.3849	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-46}	2.6255E+2	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	9.3849	1.9820	2.0933	1.6497	1.6993	1.7136	1.7277	
2^{-48}	5.2510E+2	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	10.3849	1.9820	2.0933	1.6497	1.6993	1.7136	1.7276	

From table 2 it can be observed that if we use the original Shishkin mesh, then the maximum errors do not stabilize for all values of parameter N and therefore we cannot to deduce the uniform convergence.

To see the influence of the constant σ_0 in the errors associated to the global solution, we also include the numerical results for different value of this constant, for the same set of values of ε, we include tables 3 and 4; from these, we clearly see that in order to achieve the uniform convergence and the required order we need that this constant be sufficiently large. Then, in the following numerical results we always use $\sigma_0 = 5$, which also matches with our theoretical findings.

Table 3: *Maximum errors and uniform errors for the solution at midpoints on the original Shishkin mesh, for different values of σ_0 for problem (28)*

σ_0/N	16	32	64	128	256	512	1024	2048
$\sigma_0 = 1$	4.5451E+4	9.1987E+3	2.0533E+3	5.1210E+2	1.2798E+2	3.1995E+1	7.9990E+0	1.9997E+0
	2.3048	2.1635	2.0034	2.0005	2.0000	2.0000	2.0000	
$\sigma_0 = 2$	1.8386E+3	2.3801E+2	3.0764E+1	3.9274E+0	4.9618E-1	6.2313E-2	7.8035E-3	9.7607E-4
	2.9495	2.9517	2.9696	2.9846	2.9933	2.9973	2.9991	
$\sigma_0 = 3$	2.3756E+1	4.8099E+0	4.0619E-1	2.8707E-2	2.5889E-3	8.1572E-4	2.5047E-4	7.5700E-5
	2.3042	3.5658	3.8227	3.4710	1.6662	1.7034	1.7263	
$\sigma_0 = 4$	2.0336E+2	2.1716E-1	5.3665E-2	1.5293E-2	4.7000E-3	1.4505E-3	4.4621E-4	1.3466E-4
	9.8711	2.0167	1.8111	1.7021	1.6961	1.7008	1.7284	
$\sigma_0 = 5$	5.2510E+2	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	10.3849	1.9820	2.0933	1.6497	1.6993	1.7136	1.7276	
$\sigma_0 = 6$	7.6002E+2	6.3013E-1	1.6344E-1	3.6981E-2	1.1110E-2	3.3182E-3	1.0050E-3	3.0300E-4
	10.2362	1.9468	2.1440	1.7350	1.7433	1.7232	1.7298	
$\sigma_0 = 7$	8.9238E+2	9.3128E-1	2.4748E-1	5.6829E-2	1.5293E-2	4.5561E-3	1.3736E-3	4.1322E-4
	9.9042	1.9119	2.1226	1.8938	1.7470	1.7299	1.7329	
$\sigma_0 = 8$	9.5247E+2	1.2967E+0	3.5274E-1	8.2226E-2	1.9803E-2	5.8825E-3	1.8026E-3	5.4014E-4
	9.5207	1.8781	2.1010	2.0539	1.7512	1.7064	1.7386	

Table 4: *Maximum errors and uniform errors for the solution at midpoints on the modified Shishkin mesh, for different values of σ_0 for problem (28)*

σ_0/N	16	32	64	128	256	512	1024	2048
$\sigma_0 = 1$	3.6780E+1	1.3576E+0	1.9958E+0	1.9960E+0	1.9976E+0	1.9987E+0	1.9994E+0	1.9997E+0
	4.7598	-.5559	-.0002	-.0011	-.0008	-.0005	-.0002	
$\sigma_0 = 2$	1.7667E+0	3.7937E-2	2.9862E-2	1.5301E-2	7.7438E-3	3.8926E-3	1.9505E-3	9.7607E-4
	5.5413	0.3453	0.9646	0.9826	0.9923	0.9969	0.9988	
$\sigma_0 = 3$	3.5274E-1	9.9408E-2	2.4468E-2	8.2901E-3	2.5889E-3	8.1572E-4	2.5047E-4	7.5700E-5
	1.8272	2.0225	1.5614	1.6790	1.6662	1.7034	1.7263	
$\sigma_0 = 4$	7.4289E-1	2.1716E-1	5.3665E-2	1.5293E-2	4.7000E-3	1.4505E-3	4.4621E-4	1.3466E-4
	1.7744	2.0167	1.8111	1.7021	1.6961	1.7008	1.7284	
$\sigma_0 = 5$	1.2967E+0	3.9271E-1	9.9408E-2	2.3296E-2	7.4246E-3	2.2863E-3	6.9708E-4	2.1048E-4
	1.7233	1.9820	2.0933	1.6497	1.6993	1.7136	1.7276	
$\sigma_0 = 6$	2.3643E+0	6.3013E-1	1.6344E-1	3.6981E-2	1.1110E-2	3.3182E-3	1.0050E-3	3.0300E-4
	1.9077	1.9468	2.1440	1.7350	1.7433	1.7232	1.7298	
$\sigma_0 = 7$	4.0228E+0	9.3128E-1	2.4748E-1	5.6829E-2	1.5293E-2	4.5560E-3	1.3736E-3	4.1322E-4
	2.1109	1.9119	2.1226	1.8938	1.7470	1.7299	1.7329	
$\sigma_0 = 8$	5.8023E+0	1.2967E+0	3.5274E-1	8.2226E-2	1.9803E-2	5.8825E-3	1.8026E-3	5.4014E-4
	2.1618	1.8781	2.1010	2.0539	1.7512	1.7064	1.7386	

To approximate the errors for the global normalized flux, also we calculate the errors at midpoints $x = (x_i + x_{i+1})/2$, of the corresponding Shishkin mesh, which are given by

$$F_\varepsilon^N(x) = \max_x \sqrt{\varepsilon}|S'_N(x) - \widetilde{S}'_{2N}(x)|, \quad F^N = \max_\varepsilon F_\varepsilon^N.$$

From these values we obtain the rates of convergence and the ε-uniform order of convergence for the flux, by using

$$q_\varepsilon^N = \log_2\left(F_\varepsilon^N / F_\varepsilon^{2N}\right), \quad q^N = \log_2\left(F^N / F^{2N}\right).$$

Table 5 displays the obtained results; from it we deduce the almost second order of uniform convergence for the global normalized flux, in agreement with Theorem 7.

Table 5: *Maximum errors for the normalized flux at midpoints, on the modified Shishkin mesh, and rates of convergence for problem (28)*

ε/N	16	32	64	128	256	512	1024	2048
2^0	2.1511E-4	5.3969E-5	1.3847E-5	3.5087E-6	8.8323E-7	2.2158E-7	5.5486E-8	1.3896E-8
	1.9949	1.9625	1.9806	1.9901	1.9950	1.9976	1.9975	
2^{-8}	4.0077E-2	1.7304E-2	5.7603E-3	1.6693E-3	4.4991E-4	1.1683E-4	2.9769E-5	7.5137E-6
	1.2117	1.5869	1.7869	1.8915	1.9453	1.9725	1.9862	
2^{-16}	2.2594E-2	5.6359E-2	5.0841E-2	3.0713E-2	1.4391E-2	5.6732E-3	1.9930E-3	6.4929E-4
	-1.3187	0.1487	0.7271	1.0937	1.3429	1.5093	1.6180	
2^{-24}	2.3243E-2	5.6630E-2	5.0971E-2	3.0774E-2	1.4418E-2	5.6842E-3	1.9969E-3	6.5060E-4
	-1.2148	0.1519	0.7279	1.0938	1.3429	1.5092	1.6179	
2^{-32}	2.3286E-2	5.6647E-2	5.0979E-2	3.0778E-2	1.4420E-2	5.6849E-3	1.9972E-3	6.5068E-4
	-1.2825	0.1521	0.7280	1.0938	1.3429	1.5092	1.6179	
2^{-40}	2.3288E-2	5.6648E-2	5.0980E-2	3.0778E-2	1.4420E-2	5.6849E-3	1.9972E-3	6.5069E-4
	-1.2824	0.1521	0.7280	1.0938	1.3429	1.5092	1.6179	
2^{-48}	2.3289E-2	5.6648E-2	5.0980E-2	3.0778E-2	1.4420E-2	5.6849E-3	1.9972E-3	6.5069E-4
	-1.2824	0.1521	0.7280	1.0938	1.3429	1.5092	1.6179	
F^N	5.6173E-2	5.6940E-2	5.0980E-2	3.0778E-2	1.4420E-2	5.6849E-3	1.9972E-3	6.5069E-4
q^N	-.0196	0.1595	0.7280	1.0938	1.3429	1.5092	1.6179	

Again we are interested in the influence of the constant σ_0 in the errors associated to the global normalized flux; from table 6 we see that it is necessary to take appropriately this constant to achieve the required order of uniform convergence.

Table 6: *Maximum errors and uniform errors for the the normalized flux at midpoints, on the modified Shishkin mesh, for different values of σ_0 for problem (28)*

σ_0/N	16	32	64	128	256	512	1024	2048
$\sigma_0 = 1$	4.3620E-2 0.8448	2.4287E-2 0.9407	1.2653E-2 0.9761	6.4324E-3 0.9901	3.2384E-3 0.9958	1.6239E-3 0.9982	8.1299E-4 1.7359	2.4407E-4
$\sigma_0 = 2$	5.3001E-2 0.5674	3.5767E-2 0.9259	1.8826E-2 1.1974	8.2089E-3 1.3928	3.1261E-3 1.5278	1.0841E-3 1.6193	3.5287E-4 1.6815	1.1001E-4
$\sigma_0 = 3$	5.7422E-2 0.1717	5.0980E-2 0.6806	3.1806E-2 1.0360	1.5511E-2 1.2895	6.3454E-3 1.4646	2.2992E-3 1.5820	7.6798E-4 1.6601	2.4300E-4
$\sigma_0 = 4$	5.6173E-2 -.0391	5.7716E-2 0.4302	4.2836E-2 0.8808	2.3262E-2 1.1901	1.0195E-2 1.4030	3.8552E-3 1.5453	1.3209E-3 1.6389	4.2413E-4
$\sigma_0 = 5$	5.6173E-2 -.0196	5.6940E-2 0.1595	5.0980E-2 0.7280	3.0778E-2 1.0938	1.4420E-2 1.3429	5.6849E-3 1.5092	1.9972E-3 1.6179	6.5069E-4
$\sigma_0 = 6$	5.6173E-2 -.0196	5.6940E-2 0.0234	5.6024E-2 0.5735	3.7647E-2 0.9998	1.8826E-2 1.2841	7.7303E-3 1.4736	2.7835E-3 1.5971	9.2006E-4
$\sigma_0 = 7$	5.8866E-2 0.0480	5.6940E-2 -.0287	5.8086E-2 0.4128	4.3631E-2 0.9074	2.3262E-2 1.2265	9.9410E-3 1.4385	3.6677E-3 1.5765	1.2298E-3
$\sigma_0 = 8$	1.0582E-1 0.8941	5.6940E-2 -.0122	5.7422E-2 0.2404	4.8608E-2 0.8156	2.7617E-2 1.1700	1.2274E-2 1.4039	4.6383E-3 1.5561	1.5774E-3

4 Conclusions

In this paper, we have provided a cubic spline based computational technique for obtaining second order uniformly global solution and global normalized flux for singularly perturbed reaction-diffusion boundary value problems. In order to obtain uniformly global convergence theoratically, we have modified the original Shishkin mesh near transition points, which also gives better computational results as compared to results obtained on original Shishkin mesh. Finally, from the numerical experiments we observe that in order to preserve the order of uniform convergence of the numerical method, the constant used to define the transition parameter of the piecewise uniform mesh must be close to theoretically obtained minimum value.

Acknowledgements

This research was partially supported by the project MEC/FEDER MTM2007-63204 and the Diputación General de Aragón.

References

[1] R.K. Bawa and C. Clavero, *Higher Order Global Solution and Normalized flux for Singularly Perturbed Reaction-Diffusion Problems*, Submitted

[2] P.A. Farrell, A.F. Hegarty, J.J.H. Miller, E. O'Riordan, G.I. Shishkin, *Robust computational techniques for boundary layers*, Chapman & Hall/CRC Press, 2000.

[3] J.L. Gracia, F. Lisbona, and C. Clavero, *High order ε-uniform methods for singularly perturbed reaction-diffusion problems*, Lecture Notes in Computer Science **1988** (2001) 350–358.

[4] J.J.H. Miller, E. O'Riordan, and G.I. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, Singapore, 1996.

[5] S. Natesan, R.K. Bawa, C. Clavero, *An ε-Uniform Hybrid Scheme for Singularly Perturbed Reaction–Diffusion Problems*, Numerical Mathematics and Advanced Applications ENUMATH 2005 (2006) 1079–1087.

[6] S. Natesan, R.K. Bawa, C. Clavero, *A Uniformly Convergent Method For Global Solution and Normalized Flux of Singular Perturbation Problems of Reaction-Diffusion Type*, International Journal of Information and Systems Sciences **3(2)** (2007) 207-221.

[7] Mirjana Stajanovic, *Global Convergennce method for singularly perturbed boundary value problem*, J. Comp. Appl. Math. **181** (2005) 326–335.

[8] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer, Berlin, 1996.

[9] G. Sun, M. Stynes, *An almost fourth order uniformly convergent difference scheme for a semilinear singularly perturbed reaction-diffusion problem*, Numer. Math. **70** (1995) 487–500.

[10] K. Surla, *On modelling of semilinear singularly perturbed reaction-diffusion problem*, Nonlinear Analysis, Theory, Methods & Applications **30** (1997) 61–66.

[11] K. Surla and Z. Uzelac, *A uniformly accurate spline collocation method for a normalized flux*, J. Comp. Appl. Math. **166** (2004) 291–305.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008

A MLHL based model to an Agent-Based Architecture

Emilio S. Corchado[1], María A. Pellicer[1] and M. Lourdes Borrajo[2]

[1] *Dept. de Ingeniería Civil, University of Burgos, Spain*
[2] *Dept. Informática, University of Vigo*

emails: escorchado@ubu.es, lborrajo@uvigo.es

Abstract

The firms have need of a control mechanism in order to analyse whether they are achieving their goals. A tool for the decision support process has been developed based on a multi-agent system that incorporates a case-based reasoning system and automates the business control process. The case-based reasoning system automates the organization of cases and the retrieval stage by means of a Maximum Likelihood Hebbian Learning-based method, an extension of the Principal Component Analysis which groups similar cases, identifying clusters automatically in a data set in an unsupervised mode. The system has been tested in 12 small and medium companies in the textile sector, located in the northwest of Spain and the results obtained have been very encouraging.

Key words: Agents Technology; Case Based Reasoning; Maximum Likelihood Hebbian Learning

1. Introduction

All firms need to monitor their “modus operandi” and to analyse whether they are achieving their goals. As a consequence of this, it is necessary to construct models that facilitate the analysis of work carried out in changing environments, such as finance. Processes carried out inside any firm are grouped in Functions [1, 30]. A Function is a group of coordinated and related activities, which are necessary to reach the objectives of the firm and are carried out in a systematic and iterative way [2]. In turn, each one of these functions is broken down into a series of activities and each activity is composed of a number of tasks. Control procedures have to be established in the tasks to ensure that the objectives of the firm are achieved.

This paper presents a Multiagent system (MAS) which is able to analyse the activities of a firm and calculate its level of risk. The developed model is composed of four different agent types. The principal agent, whose objectives are: to identify the state or situation of each one of the activities of the company and to calculate the risk associated with this state, incorporates a case-based reasoning (CBR) system [4, 5, 6, 7]. The CBR system uses different problem solving techniques [8, 9]. Moreover, the CBR systems proposed in the framework of this research incorporate a Maximum Likelihood Hebbian Learning (MLHL) [12] based model to automate the process of case indexing and retrieval, which may be used in problems in which the cases are characterised, predominantly, by numerical information. One of the aims of this work is to improve the performance of the CBR system integrated within the principal agent by means of incorporating the MLHL into the CBR cycle stages. The ability of the Maximum Likelihood Hebbian Learning-based methods presented in this paper to cluster cases/instances and to associate cases to clusters can be used to successfully prune the case-base without losing valuable information.

This paper first presents the Maximum Likelihood Hebbian Learning based method and its theoretical background and then, the proposed multi-agent system is presented. The system results are evaluated and, finally, the conclusions are presented.

2. Maximum Likelihood Hebbian Learning Based Method

The use of Maximum Likelihood Hebbian Learning based method has been derived from the work of [13, 15, 16, 17], etc. in the field of pattern recognition as an extension of Principal Component Analysis (PCA) [10, 11]. We first review Principal Component Analysis (PCA), which has been the most frequently reported linear operation involving unsupervised learning for data compression, which aims to find that orthogonal basis which maximises the data's variance for a given dimensionality of basis. Then, the Exploratory Projection Pursuit (EPP) theory is outlined. It is shown how Maximum Likelihood Hebbian Learning based method may be derived from PCA and it could be viewed as a method of performing EPP. Finally we show why Maximum Likelihood Hebbian Learning based method is appropriated for this type of problems. This method is used by the CBR system, of one of the system agents, to index and cluster the cases and during its retrieval stage.

2.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a standard statistical technique for compressing data; it can be shown to give the best linear compression of the data in terms of least mean square error. There are several artificial neural networks which have been shown to perform PCA e.g. [10, 11]. We will apply a negative feedback implementation [18]. The basic PCA network is described by equations

(1)-(3). Let us have an N-dimensional input vector at time t, x(t), and an M-dimensional output vector, y, with W_{ij} being the weight linking input j to output i. η is a learning rate. Then, the activation passing and learning is described by Feedforward:

$$y_i = \sum_{j=1}^{N} W_{ij} x_j \ , \forall i \tag{1}$$

Feedback:

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i \tag{2}$$

Change weights:

$$\Delta W_{ij} = \eta e_j y_i \tag{3}$$

We can readily show that this algorithm is equivalent to Oja's Subspace Algorithm [10]:

$$\Delta W_{ij} = \eta e_j y_i = \eta (x_j - \sum_k W_{kj} y_k) y_i \tag{4}$$

The PCA network not only causes convergence of the weights but causes the weights to converge to span the subspace of the Principal Components of the input data. Exploratory Projection Pursuit (EPP) is a more recent statistical method aimed at solving the difficult problem of identifying structure in high dimensional data. It does this by projecting the data onto a low dimensional subspace in which we search for its structure by eye. However not all projections will reveal the data's structure equally well. We therefore define an index that measures how "interesting" a given projection is, and then represent the data in terms of projections that maximise that index. The first step in our exploratory projection pursuit is to define which indices represent interesting directions. Now "interesting" structure is usually defined with respect to the fact that most projections of high-dimensional data onto arbitrary lines through most multi-dimensional data give almost Gaussian distributions [19]. Therefore if we wish to identify "interesting" features in data, we should look for those directions onto which the data-projections are as far from the Gaussian as possible. It was shown in [20] that the use of a (non-linear) function creates an algorithm to find those values of W which maximize that function whose derivative is f() under the constraint that W is an orthonormal matrix. This was applied in [18] to the above network in the context of the network performing an Exploratory Projection Pursuit.

2.2. ε-Insensitive Hebbian Learning

It has been shown [21] that the nonlinear PCA rule

$$\Delta W_{ij} = \eta \left(x_j f(y_i) - f(y_i) \sum_k W_{kj} f(y_k) \right) \tag{5}$$

can be derived as an approximation to the best non-linear compression of the data. Thus we may start with a cost function

$$J(W) = 1^T E\left\{\left(\mathbf{x} - Wf\left(W^T\mathbf{x}\right)\right)^2\right\} \tag{6}$$

which we minimise to get the rule (5). [22] used the residual in the linear version of (6) to define a cost function of the residual

$$J = f_1(\mathbf{e}) = f_1(\mathbf{x} - W\mathbf{y}) \tag{7}$$

where $f_1 = \|.\|^2$ is the (squared) Euclidean norm in the standard linear or nonlinear PCA rule. With this choice of $f_1(\ )$, the cost function is minimised with respect to any set of samples from the data set on the assumption that the residuals are chosen independently and identically distributed from a standard Gaussian distribution. We may show that the minimisation of J is equivalent to minimising the negative log probability of the residual, $\mathbf{e}$. , if e is Gaussian.

$$\text{Let } p(\mathbf{e}) = \frac{1}{Z}\exp(-\mathbf{e}^2) \tag{8}$$

Then, we can denote a general cost function associated with this network as

$$J = -\log p(\mathbf{e}) = (\mathbf{e})^2 + K \tag{9}$$

where K is a constant. Therefore performing gradient descent on J we have

$$\Delta W \propto -\frac{\partial J}{\partial W} = -\frac{\partial J}{\partial \mathbf{e}}\frac{\partial \mathbf{e}}{\partial W} \approx \mathbf{y}(2\mathbf{e})^T \tag{10}$$

where we have discarded a less important term. See [20] for details.

In general [23], the minimisation of such a cost function may be thought to make the probability of the residuals greater dependent on the probability density function (pdf) of the residuals. Thus, if the probability density function of the residuals is known, this knowledge could be used to determine the optimal cost function. [16] investigated this with the (one dimensional) function:

$$p(\mathbf{e}) = \frac{1}{2+\varepsilon}\exp\left(-|\mathbf{e}|_\varepsilon\right) \tag{11}$$

where

$$|e|_\varepsilon \quad = \quad \begin{cases} o & \forall |e| < \varepsilon \\ |e| - \varepsilon & otherwise \end{cases} \tag{12}$$

with ε being a small scalar ≥ 0.

Fyfe and MacDonald [16] described this in terms of noise in the data set. However, we feel that it is more appropriate to state that, with this model of the pdf of the residual, the optimal $f_1(\ )$ function is the ε-insensitive cost function:

$$f_1(\mathbf{e}) = |\mathbf{e}|_\varepsilon \tag{13}$$

In the case of the negative feedback network, the learning rule is

$$\Delta W \propto -\frac{\partial J}{\partial W} = -\frac{\partial f_1(\mathbf{e})}{\partial \mathbf{e}}\frac{\partial \mathbf{e}}{\partial W} \tag{14}$$

which gives:

$$\Delta W_{ij} \quad = \quad \begin{cases} o \quad if\left|e_j\right| < \varepsilon \\ otherwise \quad \eta y\left(sign\left(e\right)\right) \end{cases} \tag{15}$$

The difference with the common Hebb learning rule is that the sign of the residual is used instead the value of the residual. Because this learning rule is insensitive to the magnitude of the input vectors x, the rule is less sensitive to outliers than the usual rule based on mean squared error. This change from viewing the difference after feedback as simply a residual rather than an error permits us to consider a family of cost functions each member of which is optimal for a particular probability density function associated with the residual.

2.3. Applying Maximum Likelihood Hebbian Learning

The Maximum Likelihood Hebbian Learning algorithm is constructed now on the bases of the previously presented concepts as outlined here. Now the ε-insensitive learning rule is clearly only one of a possible family of learning rules which are suggested by the family of exponential distributions. This family was called an exponential family in [24] though statisticians use this term for a somewhat different family. Let the residual after feedback have probability density function

$$p(\mathbf{e}) = \frac{1}{Z}\exp(-\,|\,\mathbf{e}\,|^{p}) \tag{16}$$

Then we can denote a general cost function associated with this network as

$$J = E(-\log p(\mathbf{e})) = E(|\,\mathbf{e}\,|^{p} + K) \tag{17}$$

where K is a constant independent of W and the expectation is taken over the input data set. Therefore, performing gradient descent on J, we have

$$\Delta W \propto -\frac{\partial J}{\partial W}|_{W(t-1)} = -\frac{\partial J}{\partial \mathbf{e}}\frac{\partial \mathbf{e}}{\partial W}|_{W(t-1)} \approx E\{\mathbf{y}(p\,|\,\mathbf{e}\,|^{p-1}\ sign(\mathbf{e}))^{T}\,|_{W(t-1)}\} \tag{18}$$

where T denotes the transpose of a vector and the operation of taking powers of the norm of e is on an element wise basis as it is derived from a derivative of a scalar with respect to a vector.

Computing the mean of a function of a data set (or even the sample averages) can be tedious, and we also wish to cater for the situation in which samples keep arriving as we investigate the data set and so we derive an online learning algorithm. If the conditions of stochastic approximation [25] are satisfied, we may approximate this with a difference equation. The function to be approximated is clearly sufficiently smooth and the learning rate can be made to satisfy $\eta_k \geq 0, \sum_k \eta_k = \infty, \sum_k \eta_k^2 < \infty$ and so we have the rule:

$$\Delta W_{ij} = \eta . y_i . sign\left(e_j\right) |\, e_j \,|^{p-1} \tag{19}$$

We would expect that for leptokurtotic residuals (more kurtotic than a Gaussian distribution), values of p<2 would be appropriate, while for platykurtotic residuals (less kurtotic than a Gaussian), values of p>2 would be appropriate. Researchers from the community investigating Independent Component Analysis [24, 26]

have shown that it is less important to get exactly the correct distribution when searching for a specific source than it is to get an approximately correct distribution i.e. all supergaussian signals can be retrieved using a generic leptokurtotic distribution and all subgaussian signals can be retrieved using a generic platykutotic distribution. Our experiments will tend to support this to some extent but we often find accuracy and speed of convergence are improved when we are accurate in our choice of p. Therefore the network operation is:
Feedforward:

$$y_i = \sum_{j=1}^{N} W_{ij} x_j \,, \forall_i \qquad (20)$$

Feedback:

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i \qquad (21)$$

Weights change:

$$\Delta W_{ij} = \eta . y_i . sign(e_j) | e_j |^{p-1} \qquad (22)$$

Fyfe and MacDonald [16] described their rule as performing a type of PCA, but this is not strictly true since only the original (Oja) ordinary Hebbian rule actually performs PCA. It might be more appropriate to link this family of learning rules to Principal Factor Analysis since PFA makes an assumption about the noise in a data set and then removes the assumed noise from the covariance structure of the data before performing a PCA. We are doing something similar here in that we are basing our PCA-type rule on the assumed distribution of the residual. By maximising the likelihood of the residual with respect to the actual distribution, we are matching the learning rule to the probability density function of the residual.

More importantly, we may also link the method to the standard statistical method of Exploratory Projection Pursuit: now the nature and quantification of the interestingness is in terms of how likely the residuals are under a particular model of the probability density function of the residuals.

3. Multi-agent System

This section describes the multi-agent system in detail. Although the aim is to develop a generic model useful in any type of enterprise, the initial work has focused in small to medium firms of the textile sector to facilitate the research and its evaluation. The model here presented may be extended or adapted for other sectors. Twelve companies from the North-west of Spain have collaborated in this research, working mainly for the Spanish market. After analyzing the data relative to the activities developed within a given firm, the constructed multi-agent system is able to determine the state of each of the activities and calculate the associated risk. A Firm agent has been assigned for each firm in order to collect new data and allow consults. The Expert agents help the auditors and business control experts that collaborate in the project to provide information and feedback to the multiagent system. These experts generate prototypical cases from their experience and they have help to develop the Store agent case-base.

The CBR-based agent incorporate a case-based reasoning system as reasoning mechanism. The cycle of operations of each case based reasoning system is based on the classic life cycle of a CBR system [4, 27]. This agent is communicated with the Store agent that stores the shared case base (Table 1 shows the attributes of a case). A case represents the "shape" of a given activity developed in the company.

Table 1. Case structure

PROBLEM					SOLUTION
Case number	Input vector	Function number	Activity number	Reliability	Activity State

The CBR-based agent identifies the state or situation of each of the firm's activities and calculates the risk associated with this situation. The agent uses the data for the activity, introduced by the Firm agent, to construct the problem case. For each task making up the activity analyzed, the problem case is composed of the value of the realization state for that task, and its level of importance within the activity (according to the internal auditor).

In the retrieval step, the agent communicates with the Store agent to retrieve K cases – the most similar cases to the problem case; this is done with the Maximum Likelihood Hebbian Learning proposed method. Applying equations 20 to 22 to the case base, the MLHL algorithm groups the cases in clusters automatically. The proposed indexing mechanism classifies the cases/instances automatically, clustering together those of similar structure. One of the great advantages of this technique is that it is an unsupervised method so we do not need to have any information about of the data before hand. When a new problem case is presented to the CBR system, it is identified as belonging to a particular type by applying also equations 20 to 22 to it. This mechanism may be used as an universal retrieval and indexing mechanism to be applied to any problem similar to the presented here. Maximum Likelihood Hebbian Learning techniques are used because of the size of the database and the need to group the most similar cases together in order to help retrieve the cases that most resemble the given problem.

The re-use phase aims to obtain an initial estimation of the state of the activity analysed. In order to obtain this estimation, RBF networks are used [14, 28, 29]. As in the previous phase, the number of attributes of the problem case depends on the activity analyzed. Therefore, it is necessary to establish an RBF network system, one for each of the activities to be analysed. The k cases retrieved in the previous phase are used by the RBF network as a training group that allows it to adapt its configuration to the new problem encountered before generating the initial estimation. The RBF network is characterized by its ability to adapt, to learn rapidly, and to generalize. Specifically, within this system the network acts as a mechanism capable of absorbing knowledge about a certain number of cases and generalizing from them.

The objective of the revision phase is to confirm or refute the initial solution proposed by the RBF network, thereby obtaining a final solution and calculating the control risk. In view of the initial estimation or solution generated by the RBF network, the internal auditor (through the Firm agent) will be responsible for deciding if the solution is accepted. For this it is based on the knowledge he/she retains, specifically, knowledge about the company with which he/she is working. If he/she considers that the estimation given is valid, the system will take the solution as the final solution and in the following phase of the CBR cycle, a new case will be stored in the Store agent case base consisting of the problem case and the final solution. The system will assign the case an initial reliability of 100%. If on the other hand, the internal auditor considers the solution given by the system to be invalid, he will give his own solution which the system will take as the final solution and which together with the problem case will form the new case to be stored by the Store agent in the following phase. This new case will be given a reliability of 30%. This value has been decided by various auditors which have considered to assign a reliability of 30% to the personal opinion of the internal auditor. From the final solution: state of activity, the agent calculates the control risk associated with the activity. Every activity developed in the business sector has a risk associated with it that indicates the negative influence that affects the good operation of the firm. In this study, the level of risk is valued at three levels: low, medium and high. The calculation of the level of control risk associated with an activity is based on the current state of the activity and its level of importance. This latter value was obtained after analysing data obtained from a series of questionnaires (98 in total) carried out by auditors throughout Spain. The level of control risk was then calculated from the level of importance given to the activity by the auditors and the final solution obtained after the revision phase. For this purpose, if-then rules are employed.

The last phase executed by the agent is the communication and incorporation of the system's memory managed by the Store agent of what has been learnt after resolving a new problem. Once the revision phase has been completed, after obtaining the final solution, a new case (problem + solution) is constructed, which is stored in the agent Store's memory. Apart from the overall knowledge update involving the insertion of a new case within the agent Store memory, the multi-agent system presented carries out a local adaptation of the knowledge structures that it uses. Maximum Likelihood Hebbian Learning technique contained within the prototypes related to the activity corresponding to the new case is reorganised in order to respond to the appearance of this new case, modifying its internal structure and adapting itself to the new knowledge available. In this way, the RBF network uses the new case to carry out a complete learning cycle, updating the position of its centres and modifying the value of the weightings that connect the hidden layer with the output layer.

4. Results and Conclusions

For a given company, each one of its activities was evaluated by the system, obtaining a level of risk. On the other hand, we request to six external and independent auditors that they analyzed the situation of each company. The mission of the auditors is to estimate the state of each activity, the same as the proposed system makes. Then, we compare the result of the evaluation obtained by the auditors with the result obtained by the system. The results obtained by the system are very similar to those obtained by the external auditors (Figure 1 shows the differences between the results obtained by the system and the external auditors about the function "Sales". In general, it could be said that these results demonstrate the suitability of the techniques used for their integration in the multiagent system.

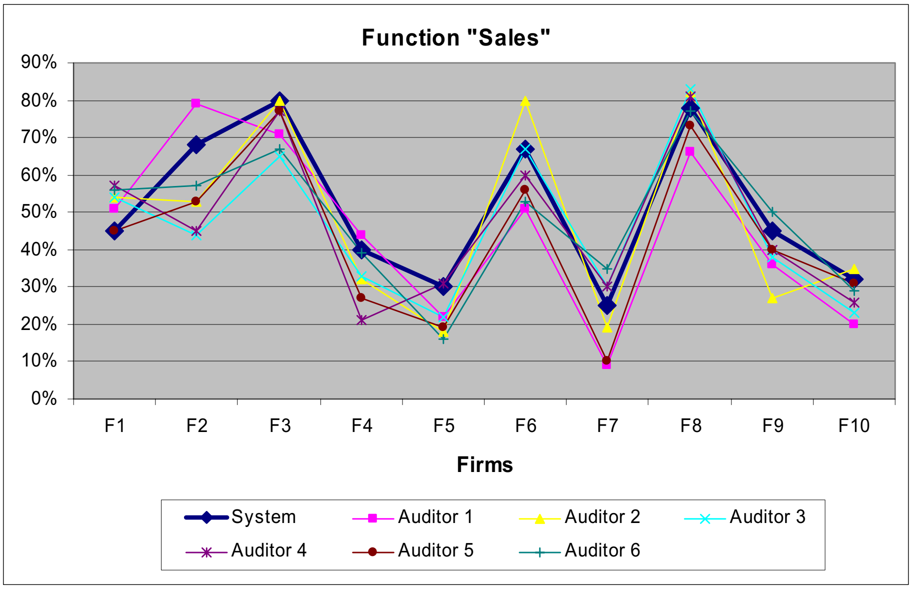


Fig. 1. Results

This article presents a multi-agent system that uses a CBR system employed as a basis for hybridization of a Maximum Likelihood Hebbian Learning technique, and a RBF net. The system is able to estimate or identify the state of the activities of the firm and their associated risk. Estimation in the environment of firms is difficult due to the complexity and the great dynamism of this environment. However, the developed model is able to estimate the state of the firm with precision. We have demonstrated a new technique for case indexing and retrieval, which could be used to construct case based reasoning systems. The basis of the method is a Maximum Likelihood Hebbian Learning algorithm. This method provides us with a very robust model for indexing the data and retrieving instances without any need of information about the structure of the data set.

5. References

[1] J.M. CORCHADO, L. BORRAJO, M.A. PELLICER AND J.C. YÁÑEZ, NEURO-SYMBOLIC *System for Business Internal Control*, IN: Lecture Notes in Computer Sciences, vol. 3275 (Springer-Verlag, Berlin, 2004) 1-10.

[2] J. MAS AND C. RAMIÓ, *La Auditoría Operativa en la Práctica* (Marcombo, Barcelona, 1997).

[3] L. BORRAJO, *Sistema híbrido inteligente aplicado a la auditoría de los sistemas internos* (Phd Thesis, Universidade de Vigo, Spain, 2003).

[4] A. AAMODT AND E. PLAZA, *Case-Based Reasoning: foundational Issues, Methodological Variations, and System Approaches*, AI Communications 7(1) (1994).

[5] J. KOLODNER, *Case-Based Reasoning* (Morgan Kaufmann, San Mateo CA, 1993).

[6] M. LENZ, B. BARTSCH-SPÖRL, D. BURKHARD AND S. WEES S. (EDS.). *Case-based Reasoning Technology: From Fundations to Applications,* IN: Lecture Notes in Artificial Intelligence 1400 (Springer-Verlag, Berlin, 1998).

[7] I. WATSON, *Applying Case-Based Reasoning: Techniques for Enterprise Systems* (Morgan Kaufmann, San Mateo CA, 1997).

[8] J. HUNT AND R. MILES, *Hybrid case-based reasoning,* The Knowledge Engineering Review, 9(4) (1994) 397-383.

[9] L.R. MEDSKER, *Hybrid Intelligent Systems* (Kluwer Academic Publishers, 1995).

[10]E. OJA, NEURAL *Networks, Principal Components and Subspaces*, International Journal of Neural Systems 1 (1989) 68-61.

[11]E. OJA, H. OGAWA AND J. WANGVIWATTANA, *Principal Components Analysis by Homogeneous Neural Networks,* part 1, The Weighted Subspace Criterion, IEICE Transaction on Information and Systems E75D (1992) 375-366.

[12]E. CORCHADO AND C. FYFE, *Maximum and Minimum Likelihood Hebbian Rules as a Exploratory Method*, IN: Proc. ICNIP'02, IEEE Catalog Number 02 EX575C. ISBN 981-04-7525-X. (SuviSoft, 2002).

[13]E. CORCHADO, D. MACDONALD AND C. FYFE, *Optimal Projections of High Dimensional Data,* in Proc. ICDM '02, The 2002 IEEE International Conference on Data Mining (IEEE Computer Society, 2002).

[14]F. FDEZ-RIVEROLA AND J.M. CORCHADO, *FSfRT: Forecasting System for Red Tides*, Applied Intelligence 21(3) (2004) 264-251.

[15]C. Fyfe and E. Corchado, *Maximum Likelihood Hebbian Rules*, in: Proc. ESANN'2002 (Bruges, 2002) 143-148.

[16]C. Fyfe and D. MacDonald, *ε-Insensitive Hebbian learning*, Neurocomputing 47(1-4) (2002) 57-35.

[17]C. Fyfe and E. Corchado, *A New Neural Implementation of Exploratory Projection Pursuit*, in: Proc. IDEAL'02, Third International Conference on Intelligent Data Engineering and Automated Learning (Manchester, 2002) 12-14.

[18]C. Fyfe and R. Baddeley, *Non-linear data structure extraction using simple Hebbian networks*, Biological Cybernetics 72(6) (1995) 541-533.

[19]P. Diaconis and D. Freedman *Asymptotics of Graphical Projections*, The Annals of Statistics 12(3) (1984) 815-793.

[20]J. Karhunen and J. Joutsensalo, *Representation and Separation of Signals Using Non-linear PCA Type Learning*, Neural Networks 7 (1994) 127-113.

[21]L. Xu, *Least Mean Square Error Reconstruction for Self-Organizing Nets*, Neural Networks 6 (1993) 648-627.

[22]P.L. Lai, D. Charles and C. Fyfe, *Seeking Independence using Biologically Inspired Artificial Neural Networks*, in: Developments in Artificial Neural Network Theory : Independent Component Analysis and Blind Source Separation (M. A. Girolami (ed.), Springer Verlag, 2000).

[23]A.J. Smola and B. Scholkopf, *A Tutorial on Support Vector Regression, Technical Report* NC2-TR-1998-030, NeuroCOLT2 Technical Report Series, 1998.

[24]A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis* (Wiley, 2002).

[25]R.L. Kashyap, C.C. Blaydon and K.S. Fu, *Stochastic Approximation, in: A Prelude to Neural Networks: Adaptive and Learning Systems* (Jerry M. Mendel (Ed), Prentice Hall, 1994).

[26]A. Hyvärinen, *Complexity Pursuit: Separating interesting components from time serie*s, Neural Computation 13 (2001) 898-883.

[27]I. Watson and F. Marir, *Case-Based Reasoning: A Review*, The Knowledge Engineering Review 9(4) (1994) 381-355.

[28]B. Fritzke, *Fast Learning with Incremental RBF Networks*, Neural Processing Letters 1(1) (1994) 5-2.

[29]J.M. Corchado, F. Díaz, L. Borrajo and F. Fdez-Riverola, *Redes Neuronales Artificiales: Un enfoque práctico* (Departamento de publicaciones de la Universidad de Vigo, Spain, 2000).

[30]J.C. Yañez. *Importancia del sistema de Control Interno en la Auditoría Legal. Contrastes Empíricos. PhD Thesis*. Universidade de Vigo (Spain). (2003)

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Fuzzy congruence relations on nd-groupoids

P. Cordero[1], **I. de las Peñas**[1], **G. Gutiérrez**[1], **J. Martínez**[1] **and M. Ojeda-Aciego**[1]

[1] *Department of Applied Mathematics, University of Málaga. Spain*

emails: pcordero@uma.es, ipcabrera@uma.es, ggutierrez@uma.es, jmartinezd@uma.es, aciego@uma.es

Abstract

In this work we introduce the notion of fuzzy congruence relation on an nd-groupoid and study conditions on the nd-groupoid which guarantee a complete lattice structure on the set of fuzzy congruence relations. The study of these conditions allowed to construct a counterexample to the statement that the set of fuzzy congruences on a hypergroupoid is a complete lattice.

Key words: Fuzzy congruence relation, nd-groupoid, multisemilattice

1 Introduction

The systematic generalization of crisp concepts to the fuzzy case has proven to be an important theoretical tool for the development of new methods of reasoning under uncertainty, imprecision and lack of information.

Regarding the generalization level, it is important to note that the definition of fuzzy sets originally presented as mappings with codomain $[0, 1]$, was soon replaced by more general structures, for instance a complete lattice, as in the L-fuzzy sets introduced by Goguen [8]. This paper continues previous work [4, 5] which is aimed at investigating L-fuzzy sets where L has the structure of a multilattice, a structure introduced in [2] and later recovered for use in other contexts, both theoretical and applied [10, 13].

Roughly speaking, a multilattice is an algebraic structure in which the restrictions imposed on a lattice, namely, the "existence of least elements in the sets of upper bounds and greatest in the sets of lower bounds" are relaxed to the "existence of minimals and maximals, respectively, in the corresponding sets of bounds". Attending to this informal description, the main difference that one notices when working with multilattices is that the operators which compute suprema and infima are no longer single-valued, since there may be several multi-suprema or multi-infima, or may be none. This immediately leads to the theory of hyperstructures, that is, algebras whose operations are set-valued.

If A is a non-empty set and H is a family of set-valued operations on A, the ordered pair (A, H) is called a hyperalgebra (or multialgebra, or polyalgebra). The study of hyperalgebras originated in 1934 when Marty introduced the so-called hypergroups in [12]. Since then, a number of papers have been published on this topic, focussing essentially on special types of hyperalgebras (such as hypergroups, hyperrings, hyperfields, vector hyperspaces, boolean hyperalgebras, ...) and guided, sometimes by purely theoretical motivations and sometimes because of their applications in other areas.

In this paper, we will focus on the most general hyperstructures, namely hypergroupoids and nd-groupoids. Our interest in these structures arises from the fact that, in a multilattice, the operators which compute the multi-suprema and multi-infima are precisely nd-groupoids or, if we have for granted that at least a multi-supremum always exists, a hypergroupoid. Actually, some of the results will be stated just in terms of multisemilattices.

Several papers have investigated the structure of the set of fuzzy congruences on different algebraic structures [1,6,7,15,17]; and in [4,5] we initiated our research in this direction. Specifically, we focused on the theory of (crisp) congruences on a multilattice and on an nd-groupoid, as a necessary step prior studying the fuzzy congruences on multilattices and the multilattice-based generalization of the concept of L-fuzzy congruence. In this paper, we study the notion of fuzzy congruence relation on nd-groupoids.

The fact that the structure of nd-groupoid is simpler than that of a multilattice does not necessarily mean that the theory is simpler as well. We will show that, in general, the set of fuzzy congruences on an nd-groupoid is not a lattice unless we assume some extra properties. This problem led us to review some related literature and, as a result, we found one counter-example in the context of congruences on a hypergroupoid.

2 Preliminaries

We can find in the literature we find the definition of a hypergroupoid as a nonempty set endowed with a hyperoperation $* : H \times H \to 2^H \smallsetminus \{\varnothing\}$. However, we are interested in a generalization of hypergroupoid that we will call non-deterministic groupoid (nd-groupoid, for short) which also considers the empty set as possible image of the hyperoperation.

Definition 2.1 *An* nd-groupoid $(A, *)$ *is defined by an nd-operation* $* : A \times A \to 2^A$ *on a nonempty set A. The* induced power groupoid *is defined as* $(2^A, *)$ *where the operation is given by* $X * Y = \{x * y \mid x \in X, y \in Y\}$ *for all* $X, Y \subseteq A$.

Notice that the definition allows the assignment of the empty set to a pair of elements, that is $a * b = \varnothing$, this mere fact, albeit simple, represents an important difference with hypergroupoids, as it will be explained later.

The following notational conventions will be used hereafter:

- We will use multiplicative notation and, thus, the symbol of the nd-operation will be omitted.

- If $a \in A$ and $X \subseteq A$, we will denote $aX = \{ax \mid x \in X\}$ and $Xa = \{xa \mid x \in X\}$. In particular, $a\varnothing = \varnothing a = \varnothing$.

- When the result of the nd-operation is a singleton, we will often omit the braces.

As stated in the introduction, our interest in extending the concept of hypergroupoid is justified by the algebraic characterization of multilattices and multisemilattices, since the operators for multi-suprema and multi-infima are both examples of nd-operators.

With this idea in mind, we introduce below the extension to the framework of nd-groupoids of some well-known properties. Assume that $(A, \cdot)$ is an nd-groupoid:

- **Idempotency**: $aa = a$ for all $a \in A$.

- **Commutativity**: $ab = ba$ for all $a, b \in A$.

- **Left m-associativity**: $(ab)c \subseteq a(bc)$ when $ab = b$, for all $a, b, c \in A$.

- **Right m-associativity**: $a(bc) \subseteq (ab)c$ when $bc = c$, for all $a, b, c \in A$.

- **m-associativity**: if it is left and right m-associative.

Note that the prefix 'm-' has its origin in the concept of multilattice.

We will focus our interest on the binary relation usually named *natural ordering*, which is defined by

$$a \leq b \text{ if and only if } ab = b$$

Although, in general, this relation is not an ordering, the properties above guarantee that the relation just defined is an ordering. Specifically, it is reflexive if the nd-groupoid is idempotent, the relation is antisymmetric if the nd-groupoid is commutative and, finally, it is transitive if the nd-groupoid is m-associative.

The two following properties of nd-groupoids have an important role in multilattice theory:

- C_1: $c \in ab$ implies that $a \leq c$ and $b \leq c$.

- C_2: $c, d \in ab$ and $c \leq d$ imply that $c = d$.

These two properties are named **comparability**. Similarly to lattice theory, we can define algebraically the concept of multisemilattice as an nd-groupoid that satisfies idempotency, commutativity, m-associativity and comparability laws. The ordered and the algebraic definitions of multisemilattice can be proved to be equivalent simply by considering $a \cdot b = \text{multisup}\{a, b\}$ and $\leq$ being the natural ordering (see [11, Theorem 2.11]).

Definition 2.2 *(Zadeh, [18]) Let A be a nonempty set. A* fuzzy relation *ρ on A is a fuzzy subset of $A \times A$ (i.e. ρ is a function from $A \times A$ to $[0, 1]$). ρ is* reflexive *in A if $\rho(x, x) = 1$ for all $x \in A$, ρ is* symmetric *in A if $\rho(x, y) = \rho(y, x)$ for all $x, y \in A$, finally, ρ is* transitive *if*

$$\sup_{z\in A} \min \{\rho(x, z), \rho(z, y)\} \leq \rho(x, y) \textit{ for all } x, y \in A$$

A fuzzy equivalence relation *is a reflexive, symmetric and transitive fuzzy relation.*

Since a fuzzy relation in a nonempty set A is a fuzzy subset of $A \times A$, we can define the inclusion, intersection and union of fuzzy relations as follows: $\rho \subseteq \sigma$ if $\rho(x, y) \leq \sigma(x, y)$ for all $x, y \in A$. $\bigcap_{i\in\Lambda} \rho_i(x, y) = \inf_{i\in\Lambda} \rho_i(x, y)$ and $\bigcup_{i\in\Lambda} \rho_i(x, y) = \sup_{i\in\Lambda} \rho_i(x, y)$ for all $x, y \in A$.

Let $FEq(A)$ be the set of fuzzy equivalence relations on a non empty set A. Murali proved in [14] that $(FEq(A), \subseteq)$ is a complete lattice where the meet is the intersection and the join is the transitive closure of the union.

The following property is used to provide characterizations of some universal properties in terms of elements; similar definitions are used in other works about fuzzy relations.

Definition 2.3 *Let A be a nonempty set and ρ a fuzzy relation on A. We say that ρ satisfies the left (resp. right)* sup property *if for all nonempty $X \subseteq A$, there exist $x_0(resp.y_0) \in X$ such that $\sup\limits_{x\in X} \rho(x, a) = \rho(x_0, a)$ (resp. $\sup\limits_{y\in X} \rho(a, y) = \rho(a, y_0)$).*

Definition 2.4 *Let ρ be a fuzzy relation on a groupoid $(G, \cdot)$; we say that ρ is* right compatible *with $\cdot$ if $\rho(ac, bc) \geq \rho(a, b)$ for all $a, b, c \in G$; similarly, ρ is said to be* left compatible *if $\rho(ca, cb) \geq \rho(a, b)$ for all $a, b, c, \in G$. A* congruence *on G is a fuzzy equivalence relation left and right compatible.*

3 Fuzzy congruence relations on nd-groupoids

Regarding the extension of the definition of congruence to the non-deterministic case, the following definition was introduced by Bakhshi and Borzooei in [1].

Definition 3.1 *Let $(A, \cdot)$ be an nd-groupoid. Then a fuzzy relation ρ on A is said to be* left (right) compatible *if for all $u \in ax$ ($u \in xa$) there exists $v \in ay$ ($v \in ya$) and for all $v \in ay$ ($v \in ya$) there exists $u \in ax$ ($u \in xa$) such that $\rho(u, v) \geq \rho(x, y)$, for all $x, y, a \in A$ and* compatible *if it is both fuzzy left and right compatible.*

This definition explicitly uses the fact that the images of the hyperoperator are nonempty. Thus, we propose an alternative definition which generalizes the previous one and adequately handles the empty images.

As a previous step to the consideration of fuzzy congruence relations on a nd-groupoid, let us note that it is possible to extend any fuzzy relation on a set A to its powerset 2^A; this construction leads to the definition of an operator $\widehat{\ }$ from the set

$FR(A)$ of fuzzy relations on A to the set $FR(2^A)$ of fuzzy relations on 2^A. Namely, given a fuzzy relation $\rho : A \times A \to [0,1]$, its *power extension* is a fuzzy relation $\widehat{\rho} : 2^A \times 2^A \to [0,1]$ defined by

$$\widehat{\rho}(X,Y) = \Big(\bigwedge_{x\in X} \bigvee_{y\in Y} \rho(x,y) \Big) \wedge \Big(\bigwedge_{y\in Y} \bigvee_{x\in X} \rho(x,y) \Big)$$

Notice that $\widehat{\rho}(\varnothing, X) = \widehat{\rho}(X,\varnothing) = 0$, for all nonempty $X \subseteq A$, $\widehat{\rho}(\varnothing,\varnothing) = 1$ and $\widehat{\rho}(\{a\},\{b\}) = \rho(a,b)$, for all $a,b \in A$.

With this power extension of a fuzzy relation, the definition of fuzzy congruence relation on an nd-groupoid $(A,\cdot)$ follows exactly the one for the deterministic case: $\widehat{\rho}(ac,bc) \geq \rho(a,b)$, for all $a,b,c \in A$. It is easy to check that a fuzzy relation that is compatible with $\cdot$ satisfies this condition but, in general, they are not equivalent as the following example shows:

Example 3.1 *Let $A = [0,1]$ be the hypergroupoid endowed with the hyperoperation $a * b := (0,1)$ and consider the fuzzy equivalence relation $\rho(a,b) = 1 - ab$. Observe that*

$$\begin{aligned}\widehat{\rho}(a*c, b*c) &= \Big(\bigwedge_{x\in(0,1)} \bigvee_{y\in(0,1)} (1-xy) \Big) \wedge \Big(\bigwedge_{y\in(0,1)} \bigvee_{x\in(0,1)} (1-xy) \Big) = \\ &= \Big(\bigwedge_{x\in(0,1)} 1 \Big) \wedge \Big(\bigwedge_{y\in(0,1)} 1 \Big) = 1 \geq \rho(a,b)\end{aligned}$$

for all $a,b,c \in A$. However, for all $x \in 0 * c$ and $y \in b * c$, we have $\rho(x,y) < \rho(0,b) = 1$ because otherwise, we would have either $x = 0$ or $y = 0$ contradicting that $x,y \in (0,1)$. Thus, ρ is not compatible with the hyperoperation $*$.

Once we have introduced the power extension of a fuzzy relation, in order to use the above condition to define the concept of fuzzy congruence relation, we study the behaviour of the operator $\widehat{\ }$ wrt the properties of reflexivity, simmetry and transitivity.

Proposition 3.2 *Let ρ be a fuzzy relation in a non-empty set A and let $\widehat{\rho}$ be its power extension as defined above. If ρ is a fuzzy equivalence relation then so is $\widehat{\rho}$.*

Summarizing the previous considerations we can state the following definition and theorem.

Definition 3.3 *A fuzzy equivalence relation ρ on an nd-groupoid $(A,\cdot)$ is said to be a* right *(resp.* left*)* congruence relation *if $\widehat{\rho}(ac,bc) \geq \rho(a,b)$ (resp. $\widehat{\rho}(ca,cb) \geq \rho(a,b)$) for all $a,b,c \in A$. A fuzzy relation is said to be a congruence relation if it is a left and right congruence relation.*

Theorem 3.4 *Let ρ be a fuzzy equivalence relation on an nd-groupoid $(A,\cdot)$. Then, ρ is a fuzzy congruence relation if and only if $\widehat{\rho}$ is a fuzzy congruence relation in the induced power groupoid $(2^A,\cdot)$.*

The sup property, which was introduced in Definition 2.3, guarantees the equivalence between our definition of fuzzy congruence relation and the one given in [1].

Lemma 3.5 *Let ρ be a fuzzy equivalence relation on an nd-groupoid $(A, \cdot)$ which satisfies sup property. Then, ρ is a fuzzy congruence relation if and only if ρ is compatible with the nd-operation.*

4 On the lattice structure of fuzzy congruence relations

In the previous section, we introduce the map $\widehat{\ }$ defined over the lattices of fuzzy equivalence relations on an nd-groupoid A and powerset 2^A. Let us now consider this map on $FCon(A)$, the subset of $FEq(A)$ given by the fuzzy congruence relations. First, notice that Theorem 3.4 guarantees that $\widehat{\ }: FCon(A) \to FCon(2^A)$ is well defined.

In the crisp case, Murali proved in [15] that the set of fuzzy congruence relations on a groupoid X is a complete sublattice of the set of all fuzzy equivalence relations. This result might suggest that the lattice structure of $FCon(2^A)$ can be reproduced on $FCon(A)$, via the map $\widehat{\ }$. However, although $\widehat{\rho}$ is injective , since $\widehat{\rho}(\{a\},\{b\}) = \rho(a,b)$, for all $a, b \in A$, it is not surjective. If it were surjective, then for all $\Theta \in FCon(2^A)$ the following equality would hold

$$\Theta(X,Y) = \Big(\bigwedge_{x\in X}\bigvee_{y\in Y}\Theta(\{x\},\{y\})\Big) \wedge \Big(\bigwedge_{y\in Y}\bigvee_{x\in X}\Theta(\{x\},\{y\})\Big)$$

but, in general, this is not the case.

Example 4.1 *Let $(A, \cdot)$ be the nd-groupoid with $A = \{a, b\}$ and $x \cdot y = \{a\}$, for all $x, y \in A$. Consider Θ the reflexive and symmetric fuzzy relation on 2^A given by $\Theta(\{a\},\{b\}) = 1; \Theta(\{a\}, A) = \Theta(\{b\}, A) = 1/2$ and $\Theta(\varnothing,\{a\}) = \Theta(\varnothing,\{b\}) = \Theta(\varnothing, A) = 0$. It is routine calculation that Θ is a congruence relation, but*

$$\Big(\bigwedge_{a\in\{a\}}\bigvee_{y\in A}\Theta(\{a\},\{y\})\Big) \wedge \Big(\bigwedge_{y\in A}\bigvee_{a\in\{a\}}\Theta(\{a\},\{y\})\Big) =$$

$$\Big(\bigvee_{y\in A}\Theta(\{a\},\{y\})\Big) \wedge \Big(\bigwedge_{y\in A}\Theta(\{a\},\{y\})\Big) = \bigwedge_{y\in A}\Theta(\{a\},\{y\}) = 1 \neq \frac{1}{2} = \Theta(\{a\}, A).$$

Under the additional assumption of commutativity with respect to the usual composition of binary relations, Bakhshi and Borzooei [1], stated that the set of all fuzzy congruence relations on a hypergrupoid $(H, \cdot)$ is a complete lattice. The following example proves that this result is not true even in the crisp case and, thus, neither in a fuzzy framework.

Example 4.2 *Let H be the set $\{a, b, c, u_0, u_1, v_0, v_1\}$ provided with a commutative hyperoperation $*$ which is defined as follows:*

$$a * a = a * b = b * b = \{a, b\}; \quad a * c = \{u_0, u_1\};$$

$$b * c = \{v_0, v_1\} \quad \textit{and } x * y = \{c\}, \quad \textit{elsewhere}$$

Consider $R, S : H \times H \to \{0,1\}$ two binary relations, where R is the least equivalence relation containing $\{(a,b),(u_0,v_0),(u_1,v_1)\}$ and S the least equivalence relation containing $\{(a,b),(u_0,v_1),(u_1,v_0)\}$. A tedious check shows that R and S commute and are compatible with the hyperoperation $*$ (they are congruence relations). However, the intersection $R \cap S$ is not a congruence relation.

As a result of the previous example, the rest of the paper studies conditions that must be satisfied by the nd-groupoid in order to guarantee that $(FCon(A), \subseteq)$ is a lattice.

Theorem 4.1 *Let $(A, \cdot)$ be an nd-groupoid satisfying idempotency and property C_1, and let ρ be a fuzzy equivalence relation satisfying the supremum property. Then ρ is a congruence relation if and only if the following holds:*

$$\text{For all } a,b,c \in A \text{ with } a \le b \text{ we have that } \widehat{\rho}(ac, bc) \ge \rho(a,b).$$

From now on we focus on the search of properties that ensure the condition of the previous theorem.

Proposition 4.2 *Let $(A, \cdot)$ be an m-associative nd-groupoid that satisfies C_1 and, for $a,b,c \in A$, consider $a \le b$ and $z \in bc$:*

1. *There exists $w \in ac$ such that $w \le z$.*
2. *Furthermore, if $(A, \cdot)$ is commutative and C_2 holds and ρ is a fuzzy congruence relation in A, then every element w as in the previous item satisfies that $\rho(w,z) \ge \rho(a,b)$.*

In order to obtain the converse result, we need to introduce the following definition.

Definition 4.3 *An nd-operation $\cdot$ in a set A is said to be **m-distributive** when, for all $a,b,c \in A$, if $a \le b$ and $w \in ac$, then $bw \cap bc \neq \varnothing$.*

The justification of this name is that a multilattice $(A, \vee, \wedge)$ in which both operations are m-distributive satisfies the following property: for all $a, b \in A$ with $a \le b$ and $c \in A$:

1. $(a \wedge b) \vee c \subseteq (a \vee c) \wedge (b \vee c)$
2. $(a \vee b) \wedge c \subseteq (a \wedge c) \vee (b \wedge c)$

Proposition 4.4 *Let $(M, \cdot)$ be an m-distributive nd-groupoid that satisfies C_1 and $a,b,c \in M$. If $a \le b$ and $w \in ac$ then there exists $z \in bc$ such that $w \le z$.*

Notice that the properties required as hypotheses of Proposition 4.4 and Proposition 4.2 are those of a multisemilattice without idempotency. The following result, stated in terms of a multisemilattice, is a straightforward consequence of these two propositions.

Proposition 4.5 *Let $(M, \cdot)$ be an m-distributive multisemilattice, ρ be a fuzzy congruence relation and $a, b, c \in M$. If $a \leq b$, $w \in ac$ and $z \in bc$ with $w \leq z$ then $\rho(w, z) \geq \rho(a, b)$.*

Now, we have all the required properties and lemmas needed in order to face the main goal of this paper, namely, to prove that under certain circumstances the set of congruences of an nd-groupoid is a complete lattice.

Theorem 4.6 *The set of the fuzzy congruence relations in an m-distributive multisemilattice M, $FCon(M)$, is a sublattice of $FEq(M)$ and, moreover is a complete lattice wrt the fuzzy inclusion ordering.*

5 Conclusions and future work

Starting with the usual notion of fuzzy congruence relation in a groupoid, we have introduced the definition of fuzzy congruence relation in an nd-groupoid by means of the power extension of the relation to the power set of the carrier. Our definition is proved to be an adequate generalization of that introduced by Bakhshi and Borzooei in [1]. Moreover, contrariwise to their claim, we have proved that, if $(A, \cdot)$ is a hypergroupoid (and thus an nd-groupoid), in general, $(FCon(A), \subseteq)$ is not a lattice.

Finally, we introduce conditions on the nd-groupoid so that we can guarantee the structure of lattice, moreover, of complete lattice of its set of fuzzy congruences. Such conditions are those of an m-distributive multisemilattice.

As future work on this research line, our plan is to keep investigating new or analogue results concerning congruences on generalized algebraic structures, specially in a non-deterministic sense; in this topic, it seems to be important to study the so-called power structures from a universal standpoint [3, 9]. We will also focus on the corresponding fuzzifications of concepts such as ideal, closure systems and homomorphisms over nd-structures, in the line of [16].

Acknowledgements

This paper has been partially supported by projects TIN2006-15455-C03-01, TIN2007-65819 (Science Ministry of Spain) and P06-FQM-02049 (Junta de Andalucía).

References

[1] M. Bakhshi and R. A. Borzooei. Lattice structure on fuzzy congruence relations of a hypergroupoid. *Inform. Sci.*, 177(16):3305–3313, 2007.

[2] M. Benado. Les ensembles partiellement ordonnés et le théorème de raffinement de Schreier. I. *Čehoslovack. Mat. Ž.*, 4(79):105–129, 1954.

[3] I. Bošnjak and R. Madarász. On power structures. *Algebra Discrete Math.*, (2):14–35, 2003.

[4] P. Cordero, G. Gutiérrez, J. Martínez, M. Ojeda-Aciego, and I. de las Peñas. Congruence relations on hypergroupoids and nd-groupoids. In *14th Spanish Conference on Fuzzy Logic and Technology*, 2008. (To appear).

[5] P. Cordero, G. Gutiérrez, J. Martínez, M. Ojeda-Aciego, and I. de las Peñas. Congruence relations on multilattices. In *8th International FLINS Conference on Computational Intelligence in Decision and Control*, 2008. (To appear).

[6] P. Das. Lattice of fuzzy congruences in inverse semigroups. *Fuzzy Sets and Systems*, 91(3):399–408, 1997.

[7] T. Dutta and B. Biswas. On fuzzy congruence of a near-ring module. *Fuzzy Sets and Systems*, 112(2):399–408, 2000.

[8] J. Goguen. L-fuzzy sets. *J. Math. Anal. Appl.*, 18:145–174, 1967.

[9] R. S. Madarász. Remarks on power structures. *Algebra Universalis*, 34(2):179–184, 1995.

[10] J. Martínez, G. Gutiérrez, I. P. de Guzmán, and P. Cordero. Generalizations of lattices via non-deterministic operators. *Discrete Math.*, 295(1-3):107–141, 2005.

[11] J. Martínez, G. Gutiérrez, I. P. de Guzmán, and P. Cordero. Multilattices via multisemilattices. In *Topics in applied and theoretical mathematics and computer science*, pages 238–248. WSEAS, 2001.

[12] F. Marty. Sur une généralisation de la notion de groupe. In *8th Congress Math. Scandinaves*, pages 45–49, Stockholm, 1934.

[13] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. Fuzzy logic programming via multilattices. *Fuzzy Sets and Systems*, 158(6):674–688, 2007.

[14] V. Murali. Fuzzy equivalence relations. *Fuzzy Sets and Systems*, 30(2):155–163, 1989.

[15] V. Murali. Fuzzy congruence relations. *Fuzzy Sets and Systems*, 41(3):359–369, 1991.

[16] U. M. Swamy and D. V. Raju. Fuzzy ideals and congruences of lattices. *Fuzzy Sets and Systems*, 95(2):249–253, 1998.

[17] Y. Tan. Fuzzy congruences on a regular semigroup. *Fuzzy Sets and Systems*, 117(3):399–408, 2001.

[18] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Derivative-free optimization and Filter Methods for solve Nonlinear Constrained Problems

Aldina Correia[1], **João Matias**[2] **and Carlos Serôdio**[3]

[1] *Escola Superior de Estudos Industriais e de Gestão, Instituto Politécnico do Porto*

[2] *Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro*

[3] *Departamento de Engenharias, Universidade de Trás-os-Montes e Alto Douro*

emails: `aldinacorreia@eseig.ipp.pt`, `j_martias@utad.pt`, `cserodio@utad.pt`

Abstract

In real problems of optimization usually the analytical expression of the objective function is not known, nor its derivatives, or are complex. In these cases one becomes essential to use optimization methods where the calculation of the derivatives, or the verification of their existence, is not necessary: the Direct Search Methods or Derivative-free Methods.

When the problem has constraints is, many times, used the penalty functions. Unfortunately the choice of the penalty parameters are, frequently, very difficult, because the most of the strategies for choosing are heuristics strategies. As an alternative to penalty function appeared the filter methods. A filter algorithm introduces a function that aggregates the constrained violations and constructs a biobjective problem. In this problem the step is accepted if it either reduces the objective function or the constrained violation. This implies the filter methods are less parameter dependent than a penalty function.

In this work we present a new direct search method, based on simplex methods, for general constrained optimization that combines the features of the simplex method and filter methods. This method does not compute or approximate any derivatives, penalty constants or Lagrange multipliers. The basics idea of simplex filter algorithm is construct a initial simplex and use the simplex to drive the search. Initially we use three of the the four basic operations of Nelder and Mead method: Reflection; Expansion and Contract. Each operation produce a new vertex that can be a good (which are a unfiltered trial point) or bad (which are filtered and leave the filter unmodified) vertex in the simplex. If the simplex search produces an unfiltered iterate, the iteration are accepted, otherwise we implement the shrink step.

Key words: Nonlinear Constrained Optimization, Filter Methods, Direct Search Methods
MSC 2000: 80M50, 49M37; 90C30

1 Introduction

We consider the general problem of nonlinear constrained optimization (NLP) of the form:

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{minimize} & f(x) \\ \text{subject to} & C(x) \geq 0 \end{array} \tag{1}$$

where $f:\mathbb{R}^n \rightarrow \mathbb{R}\cup\{\infty\}$ is the objective function and $c:\mathbb{R}^n \rightarrow (\mathbb{R}\cup\{\infty\})^m$ are the constraint functions with $C=(c_1,...,c_m)^T$. The feasible region is denoted by Ω.

Traditionally, this kind of problems is solved using penalty or merit functions that are a linear combination of the objective function and a measure of the constrained violation.

In recent years, there has been a resurgence of interest in penalty methods, mainly for exact penalty methods, [2], Byrd et al., [3], Chen et al., [7], Fletcher et al., [10], Gould et al., [14], Leyffer et al., [16], Mongeau et al., and [20], Zaslavski, because their ability to handle degenerate problems and inconsistent constrained linerizations.

The Penalty methods are designed to solve this problem by instead solving a sequence of constructed unconstrained problems, so they were seen as vehicle for solving constrained optimization problems by means of unconstrained optimization techniques. Unfortunately the choice of suitable penalty parameters are, frequently, very difficult, because the most of the strategies for choosing are heuristics strategies.

As an alternative to penalty function appeared the filter methods which are introduced by Fletcher and Leyffer [5]. Since then, the filter technique has been mostly applied, to SLP (Sequential Linear Programming) and SQP (Sequential Quadratic Programming) type methods. A filter algorithm introduces a function that aggregates constrained violations and construct a biobjective problem. In this problem the step is accepted if it either reduces the objective function or the constrained violation. This implies the filter methods are less parameter dependent than a penalty function.

The SQP-filter approach was also applied to interior point algorithms by Ulbrich et al., [19]. Audet and Dennis [1] present a pattern search filter method for derivative-free nonlinear programming. Gould et al.,[9], introduce a multidimensional filter algorithm for solving nonlinear feasibility problems. Gould et al., [11], extend the multidimensional filter techniques to general unconstrained optimization problems. Filter methods were also used in the context of nonsmooth optimization by Fletcher and Leyffer, [7] and by Karas et al. [13].

A review of filter methods is presented by Fletcher, Leyffer and Toint in [8].

Global convergence for filter in SLP problems was obtained by Fletcher, Leyffer and Toint [6]. A proof of convergence for SQP was given by Fletcher et al. in [4]. In both cases, convergence is to a point satisfying Fritz John optimally conditions. Thus, previous filter algorithms require explicit use of the derivatives of both the objective and the objective constrains.

Audet and Dennis [1] present a pattern search filter method for derivative-free nonlinear programming.

Similar to all the derivative free algorithms, the pattern search methods are suitable when some of the functions defining the problem are given as black boxes that do not assure enough precision to approximate derivatives.

These problems occur frequently in science and engineering since evaluation of the objective function usually requires a complex deterministic simulation because that describes the underlying physical phenomena. The computational noise associated with these complex simulations means that obtaining derivatives is difficult and unreliable.

Direct search methods are nonlinear optimization methods that neither require nor explicitly approximate derivatives for the problem to be solved. Instead, at each iteration a set of trial points is generated and their function values are compared with the best solution previously obtained. This information is then used to determine the next set of trial points.

In [1] Audet and Dennis don't prove such strong results, for general constraints, but they not compute or approximate any derivatives. They present and analysis a pattern search method for general constrained optimization based on filter methods for step acceptance.

The filter in [1] differs in three important aspects from the filters described above:

- it requires only simple decrease similar to unconstrained pattern-search algorithms,
- the incumbent (POLLcenter) is either feasible or the least infeasible iterate, and
- the filter includes an entry $(0, f_F)$ corresponding to a feasible iterate. A new point x_k^+ is acceptable if either of the following two conditions hold:

$$\begin{gathered} h(x_k^+) = 0 \text{ and } f(x_k^+) < f_F \\ \text{or} \\ h(x_k^+) < h_l \text{ and } f(x_k^+) < f_l, \forall l \in \mathcal{F}_k \end{gathered}$$

The authors extend the usual patter-search convergence results to filter methods.

Our goal is apply the method, of Audet and Dennis, to the other direct search methods.

The paper is divided as follow. In Section 2 we present a brief description of direct search methods. In Section 3 we briefly present the filter method of Audet and Dennis in [1]and their notation. Then, in Section 4 we present a new simplex method for general constrained optimization that combines the features of the simplex method and filter methods. This method does not compute or approximate any derivatives, penalty constants or Lagrange multipliers.

2 Direct search methods

Last decade derivative-free method or direct search methods have attracted more attention from optimization community.

These methods are especially effective when Newton-like are inappropriate or inapplicable. They are effective for minimization of a function with one or more of the fallowing properties:

- Calculation of is very expensive or time consuming
- Exact first partial derivatives of cannot be calculated
- Numerical approximation of the gradients of is impractically expensive or slow
- The values of are ”noisy”

We can organize, in accordance with Lewis et al. [15], the more popular direct search methods for unconstrained minimization into tree basic categories:

- Pattern search methods
- Simplex methods
- Methods with adaptive sets of search directions

The categorie of Pattern search methods includes methods like of Hooke and Jeeves [12] method, Nelder and Mead [17] is a good example of a Simplex method and the Powell method, [18], is a method with adaptive sets of search directions.

In this paper we just refer the Pattern search methods and the Simplex methods.

2.1 Pattern search methods

Pattern search methods are characterized by a series of exploratory moves that consider the performance of the objective function at a pattern points, all of which lie on a rational lattice.

The exploratory moves consist of a systematic strategy of visiting the points in the lattice in the instant neighborhood of the current iterate.

The Hooke and Jeeves Method, introduced in 1961 by Hooke and Jeeves [12], is one of the first direct search methods.

The algorithm consists in choose an initial point, a step-length, for the respective variables, after f is evaluated in the initial point and the method proceeds by a sequence of *exploratory* and *pattern* moves.

If an exploratory move leads to a decrease in the value of f it is called a *success*; otherwise it is called *failure*. A pattern move is not tested for success or failure.

The aim of an exploratory move is to acquire information about the function f in the neighborhood of the current base point and to find a descent direction. A Pattern move attempts to speed up the search by using information already acquired about f. It is invariably followed by a sequence of exploratory moves, with a view to finding an improved direction of search in which to make another pattern move.

2.2 Simplex Method

Simplex methods are characterized by the simple device that they use to guide the search. The basic idea of simplex search is construct a nondegenerate simplex in $\mathbb{R}^n$ and use the simplex to drive the search.

A simplex is a set of $n+1$ points in the $\mathbb{R}^n$. Thus in $\mathbb{R}^2$, a simplex is a triangle, and in $\mathbb{R}^3$ is a tetrahedron, etc.

The first know simplex method is due to Spendley in 1962. In this simplex a single move specified is the reflection, replacing a vertex by reflecting it through the centroid of the opposite face, resulting also a simplex. This move identifies the "worst" vertex in the simplex (which are the least desirable objective value) and then reflects the vertex. If the reflected vertex is still the worst vertex, then next chose the "second worst" vertex and repeat the process.

The contribution of Nelder and Mead, [17] was to turn simplex search into an optimization algorithm with additional moves designed to accelerate the search.

The four basic operations added are: Reflect; Expand; Contract and Shrink.

First they added expansion and contraction moves.

The expansion step allows for a more aggressive move by doubling the length of the step from the centroid to the reflection point, whereas the contraction steps allow for more conservative moves by halving the length of the step from the centroid to either the reflection point or the worst vertex.

Later Nelder and Mead also resolved the question of what to do if none of the steps tried bring acceptable improvement by adding a shrink step. That consists in when all else fails, reduce the lengths of the edges adjacent to the current best vertex by half.

The Nelder-Mead simplex algorithm is, of all the direct search methods, is the most often found in numerical software packages because it is effective and computationally compact.

3 Filter method of Audet and Dennis

3.1 Notation and definitions

This section is based on the work of Audet and Dennis [1].

Filter methods treat the optimization problem as biobjective attempt to minimize both functions, the objective function and a continuous function h , that aggregate constraint violation function. The priority must be given to h, at least until a feasible iterate is found.

h must satisfies:

$$h(x) \geq 0 \text{ with } h(x) = 0 \text{ if and only if } x \text{ is feasible.}$$

The function h is often set to $h(x) = \|C(x)_+\|$, where $\|.\|$ is a vector norm and $C(x)_+$ is the vector of constrained violations at x, i.e, for $i = 1, 2, ..., m$,

$$C(x)_+ = \begin{cases} C_i(x) & \text{if} \quad C_i(x) > 0 \\ 0 & \quad\ \text{otherwise} \end{cases}$$

In [1] Audet and Dennis define a second constrained violation function $h_X = h + \Psi_x$, where Ψ_x is the indicator function for X,

$$\Psi_x = \begin{cases} 0 & \text{on } X \\ +\infty & \text{elsewhere} \end{cases}$$

Others definitions are needed here.

Definição 3.1 *A point $x \in \mathbb{R}^n$ is said to **dominate** $y \in \mathbb{R}^n$, writen $x \prec y$, if $f(x) \leq f(y)$ and $h_X(x) \leq h_X(y)$ with either $f(x) < f(y)$ or $h_X(x) < h_X(y)$.*

Definição 3.2 *A filter, denoted $\mathcal{F}$, is a finite set of points in the domain of f and h such that no pair of points x and y in the set have the relation $x \prec y$.*

The filter in [1] has two additional restrictions on $\mathcal{F}$:

- Set a bound on aggregate constraint violation, so that each point satisfies
- Include only infeasible points in the filter and track feasible points separately.

With these two modifications we can define:

Definição 3.3 *A point is said to be **filtered** by a filter $\mathcal{F}$ if any of the fallowing properties hold:*

- *There exists a point $y \in \mathcal{F}$ such that $y \prec x$ or $y = x$*
- *$h_X(x) \geq h_{max}$*
- *$h_X(x) = 0$ and $f(x) \geq f^F$, where f^F is the objective function value of the best feasible point found thus far.*

Definição 3.4 *The point x is said to be **unfiltered** by $\mathcal{F}$ if is not filtered by $\mathcal{F}$.*

Definição 3.5 *Thus, the set of unfiltered points, denoted by $\bar{\mathcal{F}}$, is given by*

$$\bar{\mathcal{F}} = \bigcup_{x \in \mathcal{F}} \{y : y \prec x \text{ or } y = x\} \cup \left\{y : h_X(y) = 0, f(y) \geq f^F\right\}.$$

Note that, with this notation, if a new trial point has the same function values as those of any point in the filter, then the trial point is filtered. Thus, only the first point with such values is accepted into the filter.

3.2 The Pattern search algorithm

With the notation and definitions of the previous section we can now present the pattern search algorithm of Audet and Dennis:

1. INITIALIZATION

   Let x_0 be an undominated point of a set of initial solutions. Include all these solutions in the filter $\mathcal{F}_0$, together with $h_{max} > h(x_0)$. Fix the mesh size parameter $\Delta_0 > 0$, $(\Delta_k \to 0, k \to +\infty)$ and set the iteration counter k to 0.

2. DEFINITION OF INCUMBENT SOLUTIONS

   Define (if possible)

   f_k^F: the smallest objective function value for all feasible solutions found so far

   $h_k^I > 0$: the least positive constrained violation function value found so far

   f_k^I: the smallest objective function value of the points found so far whose constraint violation function value are equal to h_k^I.

3. SEARCH AND POOL STEPS

   Perform the SEARCH and possibly he POLL step (or only part of the steps) until an unfiltered trial point x_{k+1} is found, or when it is shown that all trial points are filtered by $\mathcal{F}_k$

   - SEARCH STEP: Evaluate the function h and f on a set of trial points on the current mesh M_k (the strategy that gives the set of points is usually provided by the user)

   - POLL STEP: Evaluate the function h and f on the poll set around p_k, where p_k satisfies either $(h(p_k), f(p_k)) = (0, f_k^F)$ or $(h(p_k), f(p_k)) = (h_k^I, f_k^I)$ a set of trial points on the current mesh M_k (the strategy that gives the set of points is usually provided by the user)

4. PARAMETER UPDATE

   If the SEARCH or POLL step produced an unfiltered iterate $x_{k+1} \in \mathcal{F}_{k+1}$, then declare the iteration *successful* and update $\Delta_{k+1} \geq \Delta_k$.

   Otherwise, set $x_{k+1} = x_k$, declare the iteration *unsuccessful* and update $\Delta_{k+1} < \Delta_k$.

   Increase $k \leftarrow k + 1$ and go back the definition of the incumbents.

4 The simplex filter algorithm

In this section we present a new derivative-free algorithm for solve nonlinear constrained problems, like Audet and Dennis method, that combines the features of a derivative free method and filter methods. That one uses the pattern search derivative-free method. Our algorithm combines filter methods with a simplex method, so it consists in a simplex method for general constrained optimization. This method does not compute or approximate any derivatives, penalty constants or Lagrange multipliers.

4.1 Notation and definitions

The terminology used by Audet and Dennis differs from the usual terminology of filter methods. That is more adjusted to derivative-free methods and so more simple to understand in this context. In this work we adopted the same notation and definitions.

4.2 The simplex search algorithm

As in the simplex methods presented in Section 2.2, the basics idea of simplex filter algorithm is construct a initial simplex and use the simplex to drive the search. The difference is that now we have a constrained problem so we introduce a filter for accept or reject the step.

Initially we use, in this method, three of the the four basic operations of Nelder and Mead: Reflection; Expansion and Contraction. We call this procedure the *simplex search*. Each operation produce a new vertex that can be a *good* (which are a unfiltered trial point) or *bad* (which are filtered and leave the filter unmodified) vertex in the present simplex.

If the simplex search produces an unfiltered iterate, then the iteration are accepted and we can find the usual best solution in the actual filter, the vertex that have the smallest objective function value of the points found so far, f_k^I, whose constraint violation function value are equal to h_k^I, the least positive constrained violation function value found so far.

Otherwise we reduce the lengths of the edges adjacent to the current best vertex by half, that is, we implement the shrink step and repeat the process.

We can now present the simplex search algorithm:

---

1. INITIALIZATION

   Let x_0 be an undominated point of a set of initial solutions. Include all these solutions in the filter $\mathcal{F}_0$, together with $h_{max} > h(x_0)$. Fix the parameters s, length of the edge step, α, β, γ, the reflection, contraction and expansion parameters and set the iteration counter k to 1.

2. DEFINITION OF INCUMBENT SOLUTIONS

   Define (if possible)

   f_k^F: the smallest objective function value for all feasible solutions found so far

   $h_k^I > 0$: the least positive constrained violation function value found so far

   f_k^I: the smallest objective function value of the points found so far whose constraint violation function value are equal to h_k^I.

3. SIMPLEX SEARCH STEPS

   Perform the SIMPLEX SEARCH STEPS until an unfiltered trial point x_k is found, or when it is shown that all trial points are filtered by $\mathcal{F}_{k-1}$

   - Construct the simplex

   - Compute the function values of all vertex of the simplex

- Identify the vertex that have the worst function value , v_p

- Identify the vertex that would go substituted the worst vertex: Reflected vertex, v_r; Expanded vertex v_e or Contract vertex, v_c

4. ACEPTANCE STEP

   If the simplex search produces an unfiltered iterate $x_k \in \mathcal{F}_k$, then the iteration are *accepted* and we can find the usual best solution in the actual filter, the vertex that have the smallest objective function value of the points found so far, f_k^I, whose constraint violation function value are equal to h_k^I, the least positive constrained violation function value found so far. Repeat the process until the stop criterion pre-established is not verified.

   Otherwise, reduce the lengths of the edges adjacent to the current best vertex by half, that is, we implement the Shrink step and repeat the process.

5 Conclusions and Future Work

In this work we present a new direct search method, based on simplex methods, for general constrained optimization that combines the features of the simplex method and filter methods. This method does not compute or approximate any derivatives.

In Future we intend to create a web page with an application that can solve any constrained and unconstrained nonlinear problem. Traditionally, this kind of problems is solved using penalty or merit functions that are a linear combination of the objective function and a measure of the constrained violation. These methods are designed to solve this problem by instead solving a sequence of constructed unconstrained problems, so they were seen as vehicle for solving constrained optimization problems by means of unconstrained optimization techniques.

Here we present an alternative to this kind of methods and other way to solve constrained nonlinear problems. We pretend implement this method in Java Language. Java has a rich variety of classes that abstracts the Internet protocols like HTTP, FTP, IP, TCP-IP, SMTP, DNS etc., so we start for implementing in java language the direct search methods and next we have as objective implement the exact penalty methods and this filter method. This next step will become possible the resolution of constrained nonlinear problems without use of derivatives or approximations of them.

References

[1] C. Audet and J.E. Dennis, *A pattern search filter method for nonlinear programming without derivatives*, SIAM Journal on Optimization: **14(4)** (2004) 980–1010.

[2] R. H. Byrd, J. Nocedal and R. A. Waltz, *Steering Exact Penalty Methods for Optimization*, Technical Report, Optimization Technology Center, Northwestern University, Evanston, IL 60208, USA, 2006.

[3] L. Chen and D. Goldfarb, *Interior-point l2-penalty methods for nonlinear programming with strong global convergence properties*, Mathematical Programming, Technical report, IEOR Dept, Columbia University, New York, 2005.

[4] R. Fletcher, N. Gould, S. Leyffer, P. L. Toint and A. Wachter, *Global convergence of trust region and SQP-filter algorithms for general nonlinear programming*, SIAM Journal on Optimization: **13(3)** (2002) 635–659.

[5] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Mathematical Programming: **Ser. A, 91(2)** (2002) 239–269.

[6] R. Fletcher, S. Leyffer and P. L. Toint, *On the global convergence of an SLP-filter algorithm*, Technical Report NA/183, Dundee University, Dept. of Mathematics, 1998.

[7] R. Fletcher and S. Leyffer, *A bundle filter method for nonsmooth nonlinear optimization*, Technical Report NA/195, Dundee University, Dept. of Mathematics, 1999.

[8] R. Fletcher, S. Leyffer and P. L. Toint, *A brief history of filter method*, Technical Report ANL/MCS-P1372-0906, Argonne National Laboratory, Mathematics and Computer Science Division, 2006.

[9] N. I. M. Gould, S. Leyffer and P. L. Toint, *A multidimensional filter algorithm for nonlinear equations and nonlinear least-squares*, SIAM Journal on Optimization: **15(1)** (2005) 17–38.

[10] N. I. M. Gould, D. Orban and P. L. Toint, *An interior-point l1-penalty method for nonlinear optimization*, Technical Report RAL-TR-2003-022 Rutherford Appleton Laboratory Chilton, Oxfordshire, UK, November 2003.

[11] N. I. M. Gould, C. Sainvitu and P. L. Toint., *A filter-trust-region method for unconstrained optimization*, SIAM Journal on Optimization: **16(2)** (2006) 341–357.

[12] R. Hooke and T. Jeeves, *Direct search solution of numerical and statistical problems*, Journal of the Association for Computing Machinery: **8** (1961) 212–229.

[13] E. W. Karas, A. A. Ribeiro, C. Sagastizábal and M. Solodov, *A bundle-filter method for nonsmooth convex constrained optimization*, Mathematical Programming: (2006) To appear.

[14] S. Leyffer, G. Lpez-Calva and Nocedal, *Interior Methods for Mathematical Programs with Complementarity Constraints*, SIAM Journal on Optimization: **17(1)** (2006) 52–77.

[15] R. M. Lewis, V. Trosset and M. W. Trosset, *Direct Search Methods: Then and Now*, NASA Langley Research Center,ICASE Report N. 2000-26 - Virginia, 2000.

[16] M. Mongeau and A. Sartenaer, *Automatic decrease of the penalty parameter in exact penalty function methods*, European Journal of Operational Research: **83(3)** (1995) 686–699.

[17] J. A. Nelder and R. Mead, *A simplex method for function minimization*, The Computer Journal: **7** (1965) 308–313.

[18] M. J. D. Powell, *An efficient method for finding the minimum of a function of several variables without calculating derivatives*, The Computer Journal: **7** (1964) 155–162.

[19] M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming*, Mathematical Programming: **Ser. A, 100(2)** (2004) 379–410.

[20] A. J. Zaslavski, *A Sufficient Condition for Exact Penalty in Constrained Optimization*, SIAM Journal on Optimization: **16(1)** (2005) 250–262.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

A Three-D Filter Line Search Method within an Interior Point Framework

M. Fernanda P. Costa[1] **and Edite M.G.P. Fernandes**[2]

[1] *Mathematics for Science and Technology Department, University of Minho, 4800-058 Guimarães, Portugal*

[2] *Production and Systems Department, University of Minho, 4710-057 Braga, Portugal*

emails: mfc@mct.uminho.pt, emgpf@dps.uminho.pt

Abstract

Here we present a primal-dual interior point three-dimensional filter line search method for nonlinear programming. The three components of the filter aim to measure adequacy of feasibility, centrality and optimality of trial iterates. The algorithm also relies on a monotonic barrier parameter reduction and it includes a feasibility/centrality restoration phase. Numerical experiments with a set of well-known problems are carried out and a comparison with a previous implementation that differs on the optimality measure is presented.

Key words: Nonlinear optimization, interior point, filter method
MSC 2000: 90C51, 90C30

1 Introduction

The filter technique of Fletcher and Leyffer [6] is used to globalize the primal-dual interior point method for solving nonlinear constrained optimization problems. This technique incorporates the concept of nondominance to build a filter that is able to reject poor trial iterates and enforce global convergence from arbitrary starting points. The filter replaces the use of merit functions, avoiding therefore the update of penalty parameters that are associated with the penalization of the constraints in merit functions.

The filter technique has already been adapted to interior point methods. In Ulbrich, Ulbrich and Vicente [12], a filter trust-region strategy based on two components is proposed. The two components combine the three criteria of the first-order optimality conditions: the first component is a measure of quasi-centrality and the second is an optimality measure combining complementarity and criticality. Global convergence to first-order critical points is also proved. In [1, 14, 15, 16], a filter line search strategy that defines two components for each entry in the filter is used. The components are the barrier objective function and the constraints violation. The global convergence is analyzed in [14]. Numerical experiments with a three-dimensional filter based line search strategy are shown in [2, 3]. The three components of the filter measure feasibility, centrality and optimality and are present in the first-order KKT conditions of the barrier problem. The optimality measure relies on the norm of the gradient of the Lagrangian function. Convergence to stationary points has been proved, although convergence to a local minimizer is not guaranteed [4].

The algorithm herein presented is a primal-dual interior point method with a three-dimensional filter line search approach that considers the barrier objective function as the optimality measure. The algorithm also incorporates a restoration phase that aims to improve either feasibility or centrality. In the paper, a performance evaluation is also carried out using a benchmarking tool, known as performance profiles [5], to compare different practical details.

The paper is organized as follows. Section 2 briefly describes the interior point method and Section 3 is devoted to introduce the 3-D filter line search method. Section 4 describes the numerical experiments that were carried out in order to analyze the performance of the new algorithm and to compare its behavior with a previous implementation that differs on the optimality measure. Conclusions are made in Section 5.

2 The interior point method

The formulation of the nonlinear constrained optimization problem that is considered in the paper is the following:

$$\begin{array}{l} \min_{x \in \mathrm{I\!R}^n} \ F(x) \\ \text{s.t.} \ \ h(x) \geq 0 \end{array} \tag{1}$$

where $h_i : \mathrm{I\!R}^n \rightarrow \mathrm{I\!R}$ for $i = 1, \ldots, m$ and $F : \mathrm{I\!R}^n \rightarrow \mathrm{I\!R}$ are nonlinear and twice continuously differentiable functions.

In this interior point paradigm, problem (1) is transformed into an equality constrained problem by using nonnegative slack variables w, as follows:

$$\begin{array}{rl} \min_{x \in \mathrm{I\!R}^n, w \in \mathrm{I\!R}^m} & \varphi_\mu(x,w) \equiv F(x) - \mu \sum_{i=1}^{m} \log(w_i) \\ \text{s.t.} & h(x) - w = 0 \\ & w \geq 0, \end{array} \tag{2}$$

where $\varphi_\mu(x,w)$ is the barrier function and μ is a positive barrier parameter [11, 13]. This is the barrier problem associated with (1). Under acceptable assumptions, the sequence of solutions of the barrier problem converges to the solution of the problem (1) when $\mu \searrow 0$. Thus, primal-dual interior point methods aim to solve a sequence of barrier problems for a positive decreasing sequence of μ values. The first-order KKT conditions for a minimum of (2) define a nonlinear system of $n+2m$ equations in $n+2m$

unknowns

$$\left\{\begin{array}{l} \nabla F(x) - A^T y = 0 \\ -\mu W^{-1} e + y = 0 \\ h(x) - w = 0 \end{array}\right. \tag{3}$$

where ∇F is the gradient vector of F, A is the Jacobian matrix of the constraints h, y is the vector of dual variables, $W = diag(w_i)$ is a diagonal matrix, and e is an m vector of all ones. Applying the Newton's method to solve (3), the following system, after symmetrization, appears

$$\left[\begin{array}{ccc} -H & 0 & A^T \\ 0 & -\mu W^{-2} & -I \\ A & -I & 0 \end{array}\right] \left[\begin{array}{c} \triangle x \\ \triangle w \\ \triangle y \end{array}\right] = \left[\begin{array}{c} \sigma \\ -\gamma_\mu \\ \rho \end{array}\right] \tag{4}$$

where

$$H = \nabla^2 F\left(x\right) - \sum_{i=1}^{m} y_i \nabla^2 h_i(x)$$

is the Hessian matrix of the Lagrangian function ($\mathcal{L} = \varphi_\mu(x, w) - y^T(h(x) - w)$) and

$$\sigma = \nabla_x \mathcal{L} = \nabla F(x) - A^T y, \;\; \gamma_\mu = \mu W^{-1} e - y \;\text{ and }\; \rho = w - h(x).$$

Since the second equation in (4) can be used to eliminate Δw without producing any off-diagonal fill-in in the remaining system, one obtains

$$\Delta w = \mu^{-1} W^2 \left(\gamma_\mu - \Delta y\right), \tag{5}$$

and the resulting reduced KKT system

$$\left[\begin{array}{cc} -H & A^T \\ A & \mu^{-1} W^2 \end{array}\right] \left[\begin{array}{c} \Delta x \\ \Delta y \end{array}\right] = \left[\begin{array}{c} \sigma \\ \pi \end{array}\right] \tag{6}$$

where $\pi = \rho + \mu^{-1} W^2 \gamma_\mu$, to compute the search directions Δx, Δw, Δy. Given initial approximations to the primal, slack and dual variables x_0, $w_0 > 0$ and $y_0 > 0$, this interior point method implements a line search procedure that chooses iteratively a step size α_k, at each iteration, and defines a new approximation by

$$\begin{array}{l} x_{k+1} = x_k + \alpha_k \Delta x_k \\ w_{k+1} = w_k + \alpha_k \Delta w_k \\ y_{k+1} = y_k + \alpha_k \Delta y_k. \end{array}$$

The choice of the step size α_k is a very important issue in nonconvex optimization and in the interior point context aims:

1. to ensure the nonnegativity of the slack and dual variables;

2. to enforce progress towards feasibility, centrality and optimality.

To decide which trial step size is accepted, at each iteration, a backtracking line search framework combined with a three-D filter method is used. This is the subject of the next section.

3 Three-D filter line search method

The methodology of a filter as outline in [6] is adapted to this interior point method. We use a three-dimensional filter. In the sequel, we use the vectors:

$$u=(x,w,y),\ u^1=(x,w),\ u^2=(w,y),$$
$$\Delta=(\Delta x,\Delta w,\Delta y),\ \Delta^1=(\Delta x,\Delta w),\ \Delta^2=(\Delta w,\Delta y).$$

To define the three components of the filter, we make use of the first-order optimality conditions (3) and the barrier objective function. The first component of the filter measures feasibility, the second measures centrality and the third represents optimality, and they are defined as follows:

$$\theta_f(u^1)=\|\rho\|_2\,,\ \theta_c(u^2)=\|\gamma_\mu\|_2 \text{ and } \varphi_\mu(u^1).$$

We remark that our previous work [2, 3] considered the norm of the gradient of the Lagrangian function in the optimality measure, therein denoted by $\theta_{op}=\frac{1}{2}\|\nabla_x\mathcal{L}\|^2$. While promoting convergence to stationary points [4], the algorithm did not enforce a sufficient decrease in the barrier function. Nonetheless, the practical implementation of the algorithm has shown convergence to minimizers even when saddle points and maximizers are present.

At each iteration k, a backtracking line search framework generates a decreasing sequence of step sizes

$$\alpha_{k,l}\in(0,\alpha_k^{\max}]\,,l=0,1,...,$$

with $\lim_l \alpha_{k,l}=0$, until a set of acceptance conditions are satisfied. Here, l denotes the iteration counter for the inner loop. $\alpha_k^{\max}$ is the longest step size that can be taken along the search directions to ensure the nonnegativity condition $u_k^2\geq 0$. Assuming that the initial approximation satisfies $u_0^2>0$, the maximal step size $\alpha_k^{\max}\in(0,1]$ is defined by

$$\alpha_k^{\max}=\max\{\alpha\in(0,1]:u_k^2+\alpha\Delta_k^2\geq(1-\varepsilon)u_k^2\} \tag{7}$$

for a fixed parameter $\varepsilon\in(0,1)$.

In this interior point context, the trial iterate $u_k(\alpha_{k,l})=u_k+\alpha_{k,l}\Delta_k$ is acceptable by the filter, if it leads to sufficient progress in one of the three measures compared to the current iterate,

$$\begin{gathered}\theta_f(u_k^1(\alpha_{k,l}))\leq\left(1-\gamma_{\theta_f}\right)\theta_f(u_k^1)\quad\text{or}\quad\theta_c(u_k^2(\alpha_{k,l}))\leq(1-\gamma_{\theta_c})\,\theta_c(u_k^2)\\ \text{or}\quad\varphi_\mu(u_k^1(\alpha_{k,l}))\leq\varphi_\mu(u_k^1)-\gamma_\varphi\theta_f(u_k^1)\end{gathered} \tag{8}$$

where $\gamma_{\theta_f},\gamma_{\theta_c},\gamma_\varphi\in(0,1)$ are fixed constants. However, to prevent convergence to a feasible but nonoptimal point, and whenever for the trial step size $\alpha_{k,l}$, the following switching conditions

$$\begin{gathered}m_k(\alpha_{k,l})<0\quad\text{and}\quad[-m_k(\alpha_{k,l})]^{s_o}\,[\alpha_{k,l}]^{1-s_o}>\delta\left[\theta_f(u_k^1)\right]^{s_f}\\ \text{and}\quad[-m_k(\alpha_{k,l})]^{s_o}\,[\alpha_{k,l}]^{1-s_o}>\delta\left[\theta_c(u_k^2)\right]^{s_c}\end{gathered} \tag{9}$$

hold, with fixed constants $\delta > 0$, $s_f > 1, s_c > 1$, $s_o \geq 1$, where

$$m_k(\alpha) = \alpha \nabla \varphi_\mu(u_k^1)^T \Delta_k^1,$$

then the trial iterate must satisfy the Armijo condition

$$\varphi_\mu(u_k^1(\alpha_{k,l})) \leq \varphi_\mu(u_k^1) + \eta_o m_k(\alpha_{k,l}), \tag{10}$$

instead of (8), to be acceptable. Here, $\eta_o \in (0, 0.5)$ is a constant. A trial step size $\alpha_{k,l}$ is called a φ-step if (10) holds. Similarly, if a φ-step is accepted as the final step size α_k in iteration k, then k is referred to as a φ-type iteration (see also [14]).

To prevent cycling between iterates that improve either the feasibility, or the centrality, or the optimality, at each iteration k, the algorithm maintains a filter that is a set $\overline{F}_k$ that contains values of θ_f, θ_c and φ_μ, that are prohibited for a successful trial iterate in iteration k [12, 14, 15, 16]. Thus, a trial iterate $u_k(\alpha_{k,l})$ is acceptable, if

$$\left(\theta_f(u_k^1(\alpha_{k,l})), \theta_c(u_k^2(\alpha_{k,l})), \varphi_\mu(u_k^1(\alpha_{k,l}))\right) \notin \overline{F}_k.$$

The filter is initialized to

$$\overline{F}_0 \subseteq \left\{(\theta_f, \theta_c, \varphi_\mu) \in \mathrm{I\!R}^3 : \theta_f \geq \theta_f^{\max}, \theta_c \geq \theta_c^{\max}, \varphi_\mu \geq \varphi_\mu^{\max}\right\}, \tag{11}$$

for the nonnegative constants $\theta_f^{\max}$, $\theta_c^{\max}$ and $\varphi_\mu^{\max}$; and is updated whenever the accepted step size satisfies (8) by

$$\begin{aligned} \overline{F}_{k+1} = \overline{F}_k \cup \big\{(\theta_f, \theta_c, \varphi_\mu) \in \mathrm{I\!R}^3 :\ & \theta_f \geq \left(1 - \gamma_{\theta_f}\right) \theta_f(u_k^1) \text{ and } \theta_c \geq \left(1 - \gamma_{\theta_c}\right) \theta_c(u_k^2) \\ & \text{and } \varphi_\mu \geq \varphi_\mu(u_k^1) - \gamma_\varphi \theta_f(u_k^1)\big\}. \end{aligned} \tag{12}$$

We remark that the filter remains unchanged whenever (9) and (10) hold for the accepted step size.

Finally, when the backtracking line search cannot find a trial step size $\alpha_{k,l}$ that satisfies the above criteria, we define a minimum desired step size $\alpha_k^{\min}$, using linear models of the involved functions,

$$\alpha_k^{\min} = \gamma_\alpha \begin{cases} \min\left\{\gamma_{\theta_f}, \frac{\gamma_\varphi \theta_f(u_k^1)}{-m_k(\alpha_{k,l})}, \frac{\delta\left[\theta_f(u_k^1)\right]^{s_f}}{\left[-m_k(\alpha_{k,l})\right]^{s_o}}, \frac{\delta\left[\theta_c(u_k^2)\right]^{s_c}}{\left[-m_k(\alpha_{k,l})\right]^{s_o}}\right\}, \text{ if } m_k(\alpha_{k,l}) < 0 \\ \qquad\qquad \text{and } (\theta_f(u_k^1) \leq \theta_f^{\min} \text{ or } \theta_c(u_k^2) \leq \theta_c^{\min}) \\ \min\left\{\gamma_{\theta_f}, \frac{\gamma_\varphi \theta_f(u_k^1)}{-m_k(\alpha_{k,l})}\right\}, \text{ if } m_k(\alpha_{k,l}) < 0 \\ \qquad\qquad \text{and } (\theta_f(u_k^1) > \theta_f^{\min} \text{ and } \theta_c(u_k^2) > \theta_c^{\min}) \\ \gamma_{\theta_f}, \text{ otherwise} \end{cases} \tag{13}$$

for positive constants $\theta_f^{\min}, \theta_c^{\min}$ and a safety factor $\gamma_\alpha \in (0, 1]$. Whenever the backtracking line search finds a trial step size $\alpha_{k,l} < \alpha_k^{\min}$, the algorithm reverts to a restoration phase. Here, the algorithm tries to find a new iterate u_{k+1} that is acceptable to the current filter, *i.e.*, (8) holds, by reducing either the constraints violation or the centrality within an iterative process.

3.1 Restoration phase

The task of the restoration phase is to compute a new iterate acceptable to the filter by decreasing either the feasibility or the centrality, whenever the backtracking line search procedure cannot make sufficient progress and the step size becomes too small. Thus, the restoration algorithm works with the new functions

$$\theta_{2,f}(u^1) = \frac{1}{2}\left\|\rho\right\|_2^2 \ \text{ and } \ \theta_{2,c}(u^2) = \frac{1}{2}\left\|\gamma_\mu\right\|_2^2$$

and the steps Δ^1 and Δ^2 that are descent directions for $\theta_{2,f}(u^1)$ and $\theta_{2,c}(u^2)$, respectively (as shown in Theorem 2 below).

3.2 Descent properties

While the search directions are computed from solving the reduced KKT system (6), we need for subsequent analysis the explicit formulas for Δx and Δw. Let $N(u) = H + \mu A^T W^{-2} A$ denote the dual normal matrix.

Theorem 1 *If N is nonsingular, then (4) has a unique solution. In particular,*

$$\begin{aligned} \Delta x &= -N^{-1}\nabla F(x) + \mu N^{-1}A^T W^{-1}e + \mu N^{-1}A^T W^{-2}\rho \\ \Delta w &= -AN^{-1}\nabla F(x) + \mu AN^{-1}A^T W^{-1}e - \left(I - \mu AN^{-1}A^T W^{-2}\right)\rho. \end{aligned}$$

Proof. Solving the second block of equations in (6) for Δy and eliminating Δy from first block of equations yields a system involving only Δx whose solution is

$$\begin{aligned} \Delta x &= N^{-1}\left(-\sigma + A^T(\mu W^{-2}\rho + \gamma_\mu)\right) \\ &= N^{-1}\left(-\nabla F(x) + A^T y + A^T(\mu W^{-2}\rho + \mu W^{-1}e - y)\right) \\ &= -N^{-1}\nabla F(x) + N^{-1}A^T y + \mu N^{-1}A^T W^{-2}\rho + \mu N^{-1}A^T W^{-1}e - N^{-1}A^T y \\ &= -N^{-1}\nabla F(x) + \mu N^{-1}A^T W^{-2}\rho + \mu N^{-1}A^T W^{-1}e \end{aligned}$$

where we used the definitions of σ and γ_μ. Using this formula of Δx, we can then solve for Δy and finally for Δw.

The resulting formula for Δw is:

$$\begin{aligned} \Delta w &= \mu^{-1}W^2\left(\gamma_\mu - \Delta y\right) \\ &= \mu^{-1}W^2\gamma_\mu - \mu^{-1}W^2\left(\mu W^{-2}\rho + \gamma_\mu - \mu W^{-2}A\Delta x\right) \\ &= \mu^{-1}W^2\gamma_\mu - \rho - \mu^{-1}W^2\gamma_\mu + A\Delta x \\ &= -\rho + AN^{-1}\left(-\sigma + A^T(\mu W^{-2}\rho + \gamma_\mu)\right) \\ &= -\rho - AN^{-1}\sigma + AN^{-1}A^T(\mu W^{-2}\rho + \gamma_\mu) \\ &= -\rho + \mu AN^{-1}A^T W^{-2}\rho - AN^{-1}\sigma + AN^{-1}A^T\gamma_\mu \\ &= -\left(I - \mu AN^{-1}A^T W^{-2}\right)\rho - AN^{-1}\left(\nabla F(x) - A^T y\right) + AN^{-1}A^T\left(\mu W^{-1}e - y\right) \\ &= -AN^{-1}\nabla F(x) + \mu AN^{-1}A^T W^{-1}e - \left(I - \mu AN^{-1}A^T W^{-2}\right)\rho. \end{aligned}$$

■

Theorem 2 *The search directions have the following properties: (i) If the dual matrix N is positive definite and $\rho = 0$, then*

$$\nabla \varphi_\mu^T \Delta^1 \leq 0.$$

ii) Furthermore

$$\nabla \theta_{2,f}^T \Delta^1 \leq 0 \;\; \textit{and} \;\; \nabla \theta_{2,c}^T \Delta^2 \leq 0.$$

Proof. First we prove (i). It is easy to see that $\nabla_x \varphi_\mu = \nabla F$ and $\nabla_w \varphi_\mu = -\mu W^{-1} e$. Let $y = \mu W^{-1} e$ and $\sigma = \nabla F - A^T y$. From the expressions for Δx and Δw given in Theorem 1, and assuming that $\rho = 0$, we get

$$\begin{aligned}
\begin{pmatrix} \nabla F \\ -y \end{pmatrix}^T \begin{pmatrix} \Delta x \\ \Delta w \end{pmatrix} &= \nabla F^T \Delta x - y^T \Delta w \\
&= \nabla F^T \left(-N^{-1} \nabla F(x) + \mu N^{-1} A^T W^{-1} e \right) - \\
&\quad -y^T \left(-A N^{-1} \nabla F(x) + \mu A N^{-1} A^T W^{-1} e \right) \\
&= \nabla F^T \left(-N^{-1} \nabla F(x) + N^{-1} A^T y \right) - \\
&\quad -y^T \left(-A N^{-1} \nabla F(x) + A N^{-1} A^T y \right) \\
&= \nabla F^T \left(-N^{-1} (\nabla F(x) - A^T y) \right) - \\
&\quad -y^T A \left(-N^{-1} (\nabla F(x) - A^T y) \right)) \\
&= \nabla F^T \left(-N^{-1} \sigma \right) - y^T A (-N^{-1} \sigma) \\
&= \left(\nabla F^T - y^T A \right) \left(-N^{-1} \sigma \right) \\
&= -\sigma^T N^{-1} \sigma \leq 0,
\end{aligned}$$

which completes the proof of the first property. To prove (ii), we star by addressing the the feasibility measure $\theta_{2,f}$. It is easy to see that $\nabla_x \theta_{2,f} = -A^T \rho$ and $\nabla_w \theta_{2,f} = \rho$, and from (4) we get

$$\begin{aligned}
\begin{pmatrix} \nabla_x \theta_{2,f} \\ \nabla_w \theta_{2,f} \end{pmatrix}^T \begin{pmatrix} \Delta x \\ \Delta w \end{pmatrix} &= \begin{pmatrix} -A^T \rho \\ \rho \end{pmatrix}^T \begin{pmatrix} \Delta x \\ \Delta w \end{pmatrix} \\
&= \left(-\rho^T A \right) \Delta x + \left(\rho^T \right) \Delta w \\
&= -\rho^T \left(A \Delta x - \Delta w \right) \\
&= -\rho^T \rho \leq 0.
\end{aligned}$$

We now address the centrality measure $\theta_{2,c}$. It is easy to see that $\nabla_w \theta_{2,c} = -\mu W^{-2} \gamma_\mu$ and $\nabla_y \theta_{2,c} = -\gamma_\mu$, and from (4) we get

$$\begin{aligned}
\begin{pmatrix} \nabla_w \theta_{2,c} \\ \nabla_y \theta_{2,c} \end{pmatrix}^T \begin{pmatrix} \Delta w \\ \Delta y \end{pmatrix} &= \begin{pmatrix} -\mu W^{-2} \gamma_\mu \\ -\gamma_\mu \end{pmatrix}^T \begin{pmatrix} \Delta w \\ \Delta y \end{pmatrix} \\
&= \gamma_\mu^T (-\mu W^{-2}) \Delta x - \gamma_\mu^T \Delta y \\
&= \gamma_\mu^T \left(-\mu W^{-2} \Delta x - \Delta y \right) \\
&= \gamma_\mu^T \left(-\gamma_\mu \right) \\
&= -\gamma_\mu^T \gamma_\mu \leq 0.
\end{aligned}$$

■

3.3 The algorithm

Next, we present the proposed primal-dual interior point 3-D filter line search algorithm for solving constrained optimization problems.

Algorithm 1 (Interior Point 3-D Filter Line Search Algorithm)

1. Given*: Starting point x_0, $u_0^2 > 0$;*

   constants $\theta_f^{\max} \in (\theta_f(u_0^1), \infty]$; $\theta_f^{\min} \in (0, \theta_f(u_0^1)]$; $\theta_c^{\max} \in (\theta_c(u_0^2), \infty]$; $\theta_c^{\min} \in (0, \theta_c(u_0^2)]$; $\varphi_\mu^{\max} \in (\varphi_\mu(u_0^1), \infty]$; $\gamma_{\theta_f}, \gamma_{\theta_c}, \gamma_\varphi \in (0,1)$; $\delta > 0$; $s_f > 1$; $s_c > 1$; $s_o \geq 1$; $\eta_o, \eta_{\theta_{2,f}}, \eta_{\theta_{2,c}} \in (0, 0.5]$; $\varepsilon_{tol} \ll 1$; $\varepsilon \in (0,1)$; $\delta_\mu, \kappa_\mu \in [0,1)$; $\epsilon \in (0,1)$;

   compute $\mu_0 > 0$ using (14).

2. *Initialize the filter using (11) and set $k \leftarrow 0$.*

3. *Stop if termination criterion is satisfied (see (15)).*

4. *If $k \neq 0$ compute μ_k using (14).*

5. *Compute the search direction Δ_k from the linear system (6), and (5).*

6. *6.1 Compute the longest step size $\alpha_k^{\max}$ using (7) to ensure positivity of slack and dual variables. Set $\alpha_{k,l} = \alpha_k^{\max}$, $l \leftarrow 0$.*

   6.2 If $\alpha_{k,l} < \alpha_k^{\min}$, go to restoration phase in step 10. Otherwise, compute the trial iterate $u_k(\alpha_{k,l})$.

   6.3 If $\left(\theta_f(u_k^1(\alpha_{k,l})), \theta_c(u_k^2(\alpha_{k,l})), \varphi_\mu(u_k^1(\alpha_{k,l}))\right) \in \overline{F}_k$, reject the trial step size and go to step 6.6.

   6.4 If $\alpha_{k,l}$ is a φ-step size ((9) holds) and the Armijo condition (10) for the φ_μ function holds, accept the trial step and go to step 7.

   6.5 If (8) holds, accept the trial step and go to step 7. Otherwise go to step 6.6.

   6.6 Set $\alpha_{k,l+1} = \alpha_{k,l}/2$, $l \leftarrow l+1$, and go back to step 6.2.

7. *Set $\alpha_k \leftarrow \alpha_{k,l}$ and $u_{k+1} \leftarrow u_k(\alpha_k)$.*

8. *If k is not a φ-type iteration, augment the filter using (12). Otherwise, leave the filter unchanged.*

9. *Set $k \leftarrow k+1$ and go back to step 3.*

10. *Use the* Restoration Algorithm *to produce a point u_{k+1} that is acceptable to the filter, i.e., $\left(\theta_f(u_{k+1}^1), \theta_c(u_{k+1}^2), \varphi_\mu(u_{k+1}^1)\right) \notin \overline{F}_k$. Augment the filter using (12) and continue with the regular iteration in step 9.*

In the restoration phase, a sufficient reduction in one of the measures $\theta_{2,f}$ and $\theta_{2,c}$ is required for a trial step size to be acceptable. The Restoration Algorithm is as follows.

Algorithm 2 (Restoration Algorithm)

1. *Set* $\alpha_{k,0}^{\max} = \alpha_k^{\max}$, $u_{k,0} = u_k$, $l = 0$ *and start with step 5.*
2. *If* $u_{k,l}$ *is acceptable to the filter then set* $u_{k+1} = u_{k,l}$ *and stop.*
3. *Compute* $\Delta_{k,l}$ *from the linear system (6), and (5) (with* $u_k = u_{k,l}$*)*
4. *(Define the vectors* $\Delta_{k,l}^1$, $\Delta_{k,l}^2$ *which are used as search directions for the variables* $u_{k,l}^1$, $u_{k,l}^2$*.) Compute* $\alpha_{k,l}^{\max}$
5. *Set* $\alpha_k = \alpha_{k,l}^{\max}$.
6. *Compute the trial iterate* $u_{k,l}(\alpha_k)$,
   If $\theta_{2,f}(u_{k,l}^1(\alpha_k)) \leq \theta_{2,f}(u_{k,l}^1) + \alpha_k \eta_{\theta_{2,f}} \nabla\theta_{2,f}(u_{k,l}^1)^T \Delta_{k,l}^1$ *or*
   $\theta_{2,c}(u_{k,l}^2(\alpha_k)) \leq \theta_{2,c}(u_{k,l}^2) + \alpha_k \eta_{\theta_{2,c}} \nabla\theta_{2,c}(u_{k,l}^2)^T \Delta_{k,l}^2$
   then set $u_{k,l+1} = u_{k,l}(\alpha_k)$, $l = l + 1$, *and return to step 2. Otherwise* $\alpha_k \leftarrow \alpha_k/2$, *and repeat step 6.*

4 Numerical experiments

To analyze the performance of the proposed interior point 3-D filter line search method, as well as to compare with our previous implementation of the algorithm [2], we used 111 constrained problems from the Hock and Schittkowski test set [8]. The tests were done in double precision arithmetic with a Pentium 4. The algorithm is coded in the C programming language and includes an interface to AMPL to read the problems that are coded in the AMPL modeling language [7].

4.1 Implementation details

Next, we report some computational details that were undertaken during our numerical experimentation, such as, for example, the initialization of the variables, the barrier parameter evaluation and the termination criterion.

Initial quasi-Newton approximation Our algorithm is a quasi-Newton based method in the sense that a symmetric positive definite quasi-Newton BFGS approximation, B_k, is used to approximate the Hessian of the Lagrangian H, at each iteration k [9]. In the first iteration, we may set $B_0 = I$ or $B_0 =$ positive definite modification of $\nabla^2 F(x_0)$, depending on the characteristics of the problem to be solved.

Monotonic reduction of the barrier parameter To guarantee a positive decreasing sequence of μ values, the barrier parameter is updated by a formula that couples the theoretical requirement defined on the first-order KKT conditions (3) with a simple heuristic. Thus, μ is updated by

$$\mu_{k+1} = \max\left\{\epsilon, \min\left\{\kappa_\mu \mu_k, \delta_\mu \frac{w_{k+1}^T y_{k+1}}{m}\right\}\right\} \tag{14}$$

where the constants $\kappa_\mu, \delta_\mu \in (0,1)$ and the tolerance ϵ is used to prevent μ from becoming too small so avoiding numerical difficulties at the end of the iterative process.

Reevaluation of centrality and optimality measures in the filter We further remark that each time the barrier parameter is updated, the θ_c component, as well as the barrier objective function value, of points in the filter may be reevaluated using the new μ so that a fair comparison of the current point with points in the filter is made. In practice, only $\theta_c^{\max}$ and $\varphi_\mu^{\max}$ need to be reevaluated.

Initialization of variables Two possible ways to initialize the primal and dual variables consider:

1. the usual published initial x_0, and the dual variables are initialized to one;
2. the published x_0 to define the dual variables and modified primal variables by solving the simplified reduced system:

$$\begin{bmatrix} -(B_0 + I) & A^T(x_0) \\ A(x_0) & I \end{bmatrix} \begin{bmatrix} \tilde{x}_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} \nabla F(x_0) \\ 0 \end{bmatrix}.$$

   Further, if $\|y_0\|_\infty > 10^3$ then we set (component-wise) $y_0 = 1$. Similarly, if $\|\tilde{x}_0\|_\infty > 10^3 \|x_0\|_\infty$, we set $\tilde{x}_0 = x_0$.

The nonnegativity of the initial slack variables are ensured by computing $w_0 = \max\{|h(x_0)|, \epsilon_w\}$, for the previously defined x_0, and a fixed positive constant ϵ_w.

Termination criterion The termination criterion considers dual and primal feasibility and centrality measures

$$\max\left\{\frac{\|\sigma\|_\infty}{s}, \|\rho\|_\infty, \frac{\|\gamma_\mu\|_\infty}{s}\right\} \leq \varepsilon_{tol}, \tag{15}$$

where

$$s = \max\left\{1, 0.01 \frac{\|y\|_1}{m}\right\}$$

and $\varepsilon_{tol} > 0$ is the error tolerance.

4.2 Parameter settings

The chosen values for the parameters involved in the Algorithms 1 and 2 are: $\theta_f^{\max} = 10^4 \max\left\{1, \theta_f(u_0^1)\right\}$, $\theta_f^{\min} = 10^{-4} \max\left\{1, \theta_f(u_0^1)\right\}$, $\theta_c^{\max} = 10^4 \max\left\{1, \theta_c(u_0^2)\right\}$, $\theta_c^{\min} = 10^{-4} \max\left\{1, \theta_c(u_0^2)\right\}$, $\varphi_\mu^{\max} = 10^4 \max\left\{0, \varphi_\mu(u_0^1)\right\}$, $\gamma_{\theta_f} = \gamma_{\theta_c} = \gamma_\varphi = 10^{-5}$,

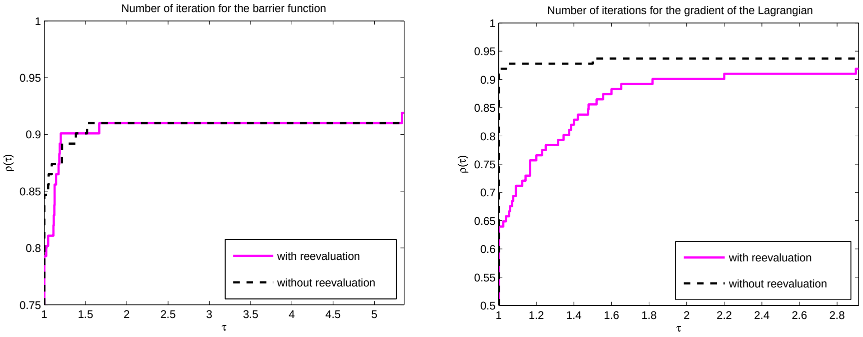


Figure 1: Profiles on the number of iterations: using φ_μ (on the left); using θ_{op} (on the right)

$\delta = 1$, $s_f = 1.1$, $s_c = 1.1$, $s_o = 2.3$, $\eta_o = \eta_{\theta_{2,f}} = \eta_{\theta_{2,c}} = 10^{-4}$, $\varepsilon = 0.95$, , $\delta_\mu = \kappa_\mu = 0.1$, $\epsilon = 10^{-9}$, $\epsilon_w = 0.01$ and $\varepsilon_{tol} = 10^{-4}$.

We carried out a set of experiments considering the two alternatives for setting the initial B_0 (as previously described) and the two ways of primal and dual variables initialization. For the subsequent analysis and comparisons we combined the overall results for each problem and selected the one which yields the smallest number of iterations.

4.3 Dolan-Moré performance profiles

To compare the performance of the reevaluation of the centrality and optimality measures in the filter for each updated μ value, we use the performance profiles as outline in [5]. These profiles represent the cumulative distribution function of a performance ratio, computed from a predefined metric. For this analysis we choose the number of iterations required to achieve the desired accuracy, as reported in (15). A brief explanation of the Dolan-Moré performance profiles follows.

Let $\mathcal{P}$ be the set of problems and $\mathcal{C}$ the set of codes used in the comparative study. Let $t_{p,c}$ be the performance metric - number of iterations required to solve problem p by code c. Then, the comparison relies on the performance ratios

$$r_{p,c} = \frac{t_{p,c}}{\min\{t_{p,c}, c \in \mathcal{C}\}}, p \in \mathcal{P}, c \in \mathcal{C}$$

and the overall assessment of the performance of a code c is given by $\rho_c(\tau) = \frac{n_{P_\tau}}{n_P}$, where n_P is the number of problems in the set $\mathcal{P}$ and n_{P_τ} is the number of problems in the set such that the performance ratio $r_{p,c}$ is less than or equal to $\tau \in \mathrm{I\!R}$ for code $c \in \mathcal{C}$. Thus, $\rho_c(\tau)$ gives the probability (for code c) that $r_{p,c}$ is within a factor τ of the best possible ratio. The function ρ_c is the cumulative distribution function for the performance ratio.

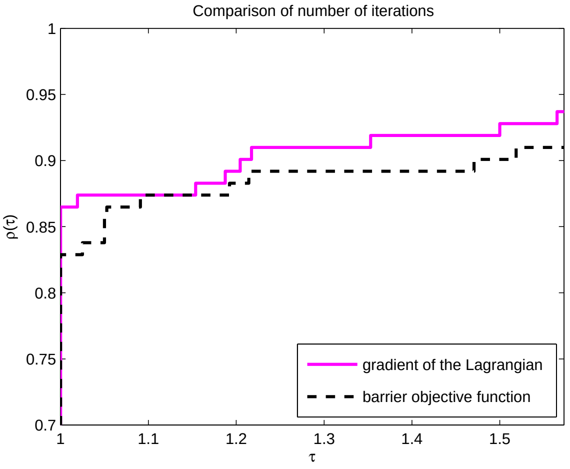


Figure 2: Profiles on the number of iterations: comparison between θ_{op} and φ_μ

First, we examine the practical performance of the reevaluation of the filter for each new μ value within the herein proposed algorithm - where φ_μ is used as the optimality measure. The performance plots on the left of Figure 1 show that the version that does not implement the reevaluation of the filter is the most efficient on 85% of the problems (see the corresponding value of $\rho(1)$). Next, the filter reevaluation process was implemented in our previous implementation of the algorithm - when θ_{op} was used as the optimality measure. The performance plots on the right of Figure 1 definitely show that the filter reevaluation yields the worst performance.

From the previous analysis, we decided to disable the reevaluation filter process from both algorithms and plot the performance profiles together. Figure 2 represents the performance profiles of the number of iterations. The use of the barrier function to measure the trial iterate optimality adequacy did not improve the performance of this interior point based method, at least when the number of iterations is the metric used in these performance profiles.

Finally, to further compare the convergence of both interior point 3-D filter line search algorithms we include Table 1 that records the objective function values at the found solutions. Only the problems that were solved at least by one of the versions in comparison are listed. While the previous implementation did not converge to the required solution on 3 problems (hs046, hs105, hs111), within 100 iterations, the new algorithm did not reach the solution on the following problems: hs064, hs083, hs101, hs106 and hs118. In all the other problems, both algorithms reach the same solution with the desired accuracy.

Table 1: Objective function values at the solution

Prob	with θ_{op}	with φ_μ	Prob	with θ_{op}	with φ_μ	Prob	with θ_{op}	with φ_μ
hs001	8.9525e-13	6.5934e-12	hs038	1.0359e-11	6.6804e-14	hs077	2.4151e-01	2.4151e-01
hs002	5.0426e-02	5.0426e-02	hs039	-1.0000e00	-1.0000e00	hs078	-2.9197e00	-2.9197e00
hs003	1.0000e-04	1.0000e-04	hs040	-2.5000e-01	-2.5000e-01	hs079	7.8777e-02	7.8777e-02
hs004	2.6667e00	2.6667e00	hs041	1.9999e00	1.9999e00	hs080	5.3949e-02	5.3949e-02
hs005	-1.9132e00	-1.9132e00	hs042	1.3858e01	1.3858e01	hs081	5.3950e-02	5.3950e-02
hs006	1.6997e-12	1.6997e-12	hs043	-4.4000e01	-4.4000e01	hs083	-3.0666e04	-
hs007	-1.7321e00	-1.7321e00	hs044	-1.5000e01	-1.5000e01	hs086	-3.2349e01	-3.2349e01
hs008	-1.0000e00	-1.0000e00	hs045	1.0000e00	1.0000e00	hs087	8.8276e03	8.8276e03
hs009	-4.9999e-01	-4.9999e-01	hs046	2.1089e-02	6.2703e-07	hs088	1.3626e00	1.3626e00
hs010	-1.0000e00	-1.0000e00	hs047	1.1658e-07	1.1658e-07	hs089	1.3626e00	1.3626e00
hs011	-8.4985e00	-8.4985e00	hs048	1.6555e-12	1.6555e-12	hs090	1.3626e00	1.3626e00
hs012	-3.0000e01	-3.0000e01	hs049	1.17559e-06	1.1756e-06	hs091	1.3626e00	1.3626e00
hs014	1.3934e00	1.3934e00	hs050	3.1993e-10	3.1993e-10	hs092	1.3627e00	1.3627e00
hs015	3.0650e02	3.0650e02	hs051	3.5090e-10	3.5090e-10	hs093	1.3508e02	1.3508e02
hs016	2.5000e-01	2.5000e-01	hs052	5.3266e00	5.3266e00	hs095	1.5620e-02	1.5620e-02
hs017	1.0000e00	1.0000e00	hs053	4.0930e00	4.0930e00	hs096	1.5620e-02	1.5620e-02
hs018	4.9999e00	5.0000e00	hs054	1.9286e-01	1.9286e-01	hs097	3.1358e00	3.1358e00
hs019	-6.9618e03	-6.9618e03	hs055	6.6667e00	6.6667e00	hs098	4.0712e00	4.0712e00
hs020	3.8199e01	3.8199e01	hs056	-1.0788e-10	-1.0788e-10	hs100	6.8063e02	6.8063e02
hs021	-9.9960e01	-9.9960e01	hs057	3.0648e-02	3.0648e-02	hs101	1.8098e03	-
hs022	1.0000e00	1.0000e00	hs059	-7.8028e00	-7.8028e00	hs102	9.1188e02	9.1188e02
hs023	2.0000e00	2.0000e00	hs060	3.2568e-02	3.2568e-02	hs103	5.4367e02	5.4367e02
hs024	-1.0000e00	-1.0000e00	hs061	-1.4365e02	-1.4365e02	hs104	3.9512e00	3.9512e00
hs025	1.8361e-10	2.7269e-11	hs062	-2.6273e04	-2.6273e04	hs105	-	1.1363e03
hs026	7.6064e-07	1.9872e-07	hs063	9.6172e02	9.6172e02	hs106	7.0492e03	-
hs027	3.9999e-02	3.9999e-02	hs064	6.2998e03	-	hs108	-5.0000e-01	-5.0000e-01
hs028	1.0270e-09	1.0270e-09	hs065	9.5354e-01	9.5354e-01	hs110	-4.5778e01	-4.5778e01
hs029	-2.2627e01	-2.2627e01	hs066	5.1816e-01	5.1816e-01	hs111	-	-4.7761e01
hs030	1.0002e00	1.0002e00	hs067	-1.1620e03	-1.1620e03	hs112	-4.7761e01	-4.7761e01
hs031	5.9999e00	6.0000e00	hs070	2.7971e-01	2.7971e-01	hs113	2.4306e01	2.4306e01
hs032	1.0000e00	1.0000e00	hs071	1.7014e01	1.7014e01	hs114	-1.7688e03	-1.7688e03
hs033	-4.5858e00	-4.5858e00	hs072	7.2760e02	7.2767e02	hs117	3.2349e01	3.2349e01
hs034	-8.3403e-01	-8.3403e-01	hs073	2.9894e01	2.9894e01	hs118	6.6482e02	-
hs035	1.1116e-01	1.1116e-01	hs074	5.1265e03	5.1265e03	hs119	2.4490e02	2.4490e02
hs036	-3.3000e03	-3.3000e03	hs075	5.1744e03	5.1744e03			
hs037	-3.4560e03	-3.4560e03	hs076	-4.6818e00	-4.6818e00			

5 Conclusions

A primal-dual interior point method based on a filter line search approach is presented. The novelty here is that each entry in the filter has three components that represent the feasibility, centrality and optimality of the iterate. Using the barrier objective function as the optimality measure, the algorithm is able to enforce a sufficient decrease of the barrier function and converge to stationary points that are minimizers. The new algorithm is tested with a set of well-known problems and compared with our previous implementation of an interior point three-dimensional filter line search [2, 3], using a benchmarking tool with performance profiles. The numerical results show that both algorithms have similar practical behaviors.

We would like to remark that the performance profiles reflect only the performance of the tested codes on the data being used. Definitive conclusions could be made if different test sets, including larger academic problems and real engineering problems [10], were used. This will be a matter of future research.

References

[1] H.Y. Benson, R.J. Vanderbei and D.F. Shanno, *Interior-point methods for nonconvex nonlinear programming: filter methods and merit functions*, Computational Optimization and Applications, **23** (2002) 257–272.

[2] M.F.P. Costa and E.M.G.P. Fernandes, *Comparison of interior point filter line search strategies for constrained optimization by performance profiles*, International Journal of Mathematics Models and Methods in Applied Sciences, **1** (2007) 111–116.

[3] M.F.P. Costa and E.M.G.P. Fernandes, *Practical implementation of an interior point nonmonotone line search filter method*, International Journal of Computer Mathematics **85** (2008) 397–409.

[4] M.F.P. Costa and E.M.G.P. Fernandes, *A globally convergent interior point filter line search method for nonlinear programming*, submitted to Journal of Numerical Analysis, Industrial and Applied Mathematics 2007.

[5] E.D. Dolan and J.J. Moré, *Benchmarking optimization software with performance porfiles*, Mathematical Programming A **91** (2002) 201–213.

[6] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Mathematical Programming **91** (2002) 239–269.

[7] R. Fourer, D.M. Gay and B. Kernighan, *A modeling language for mathematical programming*, Management Science **36** (1990) 519–554.

[8] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming*, Springer-Verlag, 1981.

[9] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer-Verlag, 1999.

[10] K. E. Parsopoulos and M. N. Vrahatis, *Unified particle swarm optimization for solving constrained engineering optimization problems*, Vol 3612 of LNCS, 582–591. Springer-Verlag, ICNC 2005 edition, 2005.

[11] D.F. Shanno and R.J. Vanderbei, *Interior-point methods for nonconvex nonlinear programming: orderings and higher-order methods*, Mathematical Programming B **87** (2000) 303–316.

[12] M. Ulbrich, S. Ulbrich and L.N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming*, Mathematical Programming **100** (2004) 379–410.

[13] R.J. Vanderbei and D.F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Computational Optimization and Applications **13** (1999) 231–252.

[14] A. Wächter and L.T. Biegler, *Line search filter methods for nonlinear programming: motivation and global convergence*, SIAM Journal on Optimization **16** (2005) 1–31.

[15] A. Wächter and L.T. Biegler, *Line search filter methods for nonlinear programming: local convergence*, SIAM Journal on Optimization **16** (2005) 32–48.

[16] A. Wächter and L.T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical Programming **106** (2007) 25–57.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Some relationships for multi-scale vulnerabilities of complex networks

R. Criado[1], J. Pello[1], M. Romance[1] and M. Vela-Pérez[1]

[1] *Departamento de Matemática Aplicada, ESCET,*
Universidad Rey Juan Carlos (Spain)

emails: `regino.criado@urjc.es`, `javier.pello@urjc.es`,
`miguel.romance@urjc.es`, `maria.vela@urjc.es`

Abstract

We show that there is a close relationship between node-based multiscale vulnerability and link-based multiscale vulnerability by giving sharp lower analytical estimates The techniques involved in the proof include estimates of the ℓ^p-norm of vectors in $\mathbb{R}^m$ and some techniques coming from the geometry of finite dimensional normed spaces.

Key words: Multi-scale vulnerability; node betweeness; link betweeness;

1 Introduction and notation

Complex systems can be used to describe many biological, social and technological systems [1, 2]. In order to understand the main structure, we have to characterise those systems by using some topological properties, among them we find the characteristic path length, clustering, efficiency or vulnerability.

In order to describe the robustness or performance of a system, we may employ a global parameter which show us the behaviour of the system under an intentional or random attack. This can be done by using a vulnerability function such as those appeared in [4], but some of them have several limitations. Hence, we provide a new parameter related to the multi-scale vulnerability for edges appeared in [3], which overcomes those limitations and can be calculated faster since the complexity of computing the new parameter is up to n times the order of complexity of computing the link-based multi-scale vulnerability. The multi-scale parameter related to the link betweenness presented in [3] has entailed a new approach to quantify the vulnerability of a complex network. If $G = (X, E)$ is a complex network with n nodes and m links, for every $p \in [1, +\infty)$, the normalised multi-scale vulnerability for edges is defined (see [3]) as

$$V_{E,p}(G) = \left(\frac{1}{m} \sum_{\ell \in E} b_\ell^p \right)^{1/p},$$

where b_ℓ is the betweeness of the link $\ell \in E$ (see, for example [6, 8]) given by

$$b_\ell = \frac{1}{n(n-1)} \sum_{\substack{j,k \in X \\ j \neq k}} \frac{n_{jk}(\ell)}{n_{jk}},$$

where $n_{jk}(\ell)$ is the number of geodesics from j to k that contain the link ℓ and n_{jk} the total number of geodesics from j to k.

In some cases, the most important ingredients of a complex network are nodes instead of links. Therefore, it is interesting to consider a normalised multi-scale vulnerability based on vertices that can be defined as

$$V_{X,p}(G) = \left(\frac{1}{n} \sum_{v \in X} b_v^p \right)^{1/p} = \left(\frac{1}{n} \sum_{v \in X} \left(\frac{1}{n(n-1)} \sum_{\substack{j,k \in X \\ j \neq k}} \frac{n_{jk}(v)}{n_{jk}} \right)^p \right)^{\frac{1}{p}},$$

where $n_{jk}(v)$ is the number of geodesics from j to k that go through v.

We will give some estimates and bounds that relate both parameters (the multi-scale vulnerability for edges and the multi-scale vulnerability for nodes) and we make a study for the Madrid underground network. We find what are the stations and routes with greatest values of these parameters, that can be understood as those stations and routes which are more important for the underground. Hence, a malfunctioning of these stations or routes will collapse the whole network.

2 Sharp estimates for node and link vulnerabilities

In this section we give a better lower bound for the multi-scales vulnerabilities for $p \neq 1$, since for $p = 1$ it was proved in [5] that there is a linear relationship between both vulnerabilities. By using some techniques from geometric convex analysis, in [5] it was proved the following result:

Theorem 2.1 ([5]) *Let $G = (X, E)$ be a network with n nodes and m links and take $1 \leq p < \infty$, then*

$$2^{\frac{1}{p}-1} \left(\frac{m}{n} \right)^{1/p} V_{E,p}(G) \leq V_{X,p}(G) \leq 2^{\frac{1}{p}-1} \left(\frac{m}{n} \right)^{1/p} (gr_{max})^{1-1/p} V_{E,p}(G) + \frac{1}{n},$$

where gr_{max} denotes the maximal degree of the network G.

The upper estimate in the previous theorem is sharp since if $G = K_n$ is the complete network with n vertices then, by easy computations we get that

$$V_{X,p}(G) = \frac{2}{n} = 2^{\frac{1}{p}-1} \left(\frac{m}{n} \right)^{1/p} (gr_{max})^{1-1/p} V_{E,p}(G) + \frac{1}{n},$$

and therefore the upper estimate is actually an equality. The lower bound is not sharp, and the main goal of this talk is to show that it can be improved by using some results coming from the geometry of $\mathbb{R}^n$. We will prove the following lower estimate:

Theorem 2.2 *Let $G=(X,E)$ be a network with n nodes and m links. If $1\leq p<\infty$, then*

$$\left[\left(\frac{1}{2}\left(\frac{2m}{n}\right)^{1/p}V_{E,p}(G)+\frac{1}{n^{1+\frac{1}{p}}}\right)^{p}+\frac{n-1}{n^{p+1}}\right]^{\frac{1}{p}}\leq V_{X,p}(G).$$

In order to prove this result, if we denote for every $\ell\in E$ and $v\in X$

$$\bar{x}_\ell=\sum_{\substack{j,k\in X\\ j\neq k}}\frac{n_{jk}(\ell)}{n_{jk}},\qquad \bar{y}_v=\sum_{\substack{j,k\in X\\ j\neq k}}\frac{n_{jk}(v)}{n_{jk}},$$

and $\bar{x}=(\bar{x}_\ell)_{\ell\in E}\in\mathbb{R}^m$, $\bar{y}=(\bar{y}_v)_{v\in X}\in\mathbb{R}^n$, then, we have that

$$\begin{aligned}(n-1)n\,m^{1/p}\,V_{E,p}(G) &= \left(\sum_{\ell\in E}\bar{x}_\ell^p\right)^{1/p}=\|\bar{x}\|_p, &(1)\\ (n-1)n^{1+1/p}\,V_{X,p}(G) &= \left(\sum_{v\in X}\bar{y}_v^p\right)^{1/p}=\|\bar{y}\|_p, &(2)\end{aligned}$$

since all $\bar{x}_\ell$ and all $\bar{y}_v$ are non-negatives. Therefore, if we want to get some estimates for $V_{E,p}(G)$ and $V_{X,p}(G)$, it is enough to relate $\|\bar{x}\|_p$ to $\|\bar{y}\|_p$. Vectors $\bar{x}$ and $\bar{y}$ have information about the geodesics that a fixed node or link contain and it was shown in [5] that there is a strong relationship between the number of geodesics that contain a node v and the number of geodesics that contain some links incident to v. We use this fact to relate $\bar{x}$ and $\bar{y}$, by using the following lemma:

Lemma 2.3 ([5]) *Let $G=(X,E)$ be a network with n nodes and m links. If we fix $v\in X$, then for every $j,k\in X$ $(j\neq k)$*

$$n_{jk}(v)=\frac{1}{2}\left(\sum_{\ell\ni v}n_{jk}(\ell)\right)+\frac{\delta_{jv}+\delta_{kv}}{2}\left(\sum_{\ell\ni v}n_{jk}(\ell)\right),$$

where $\delta_{ab}=1$ if $a=b$ and $\delta_{ab}=0$ otherwise and $\ell\ni v$ means that the link ℓ is incident to v.

By using this relationships, the proof of the main result reduces to compare the ℓ^p-norm of a vector $x=(x_1,\ldots,x_n)\in\mathbb{R}^n, x_i\geq 1$ with the ℓ^p-norm of the vector $\bar{x}=(x_1+1,\ldots,x_n+1)\in\mathbb{R}^n$. Once we have translate the problem to this framework we obtain the main result by using the following theorem.

Theorem 2.4 *Let $x=(x_1,\ldots,x_n)\in\mathbb{R}^n, x_i\geq 1$ and $\bar{x}=(x_1+1,\ldots,x_n+1)\in\mathbb{R}^n$. Then for every $1\leq p<\infty$*

$$\|\bar{x}\|_p^p\geq(\|x\|_p+1)^p+(n-1).$$

The proof also includes other tools, such as the following lemma, which is a classic and well-known consequence of the Hölder's inequality (see, for example [7]):

Lemma 2.5 ([7]) *If $a = (a_1 \dots a_k) \in \mathbb{R}^k$ is a vector and $1 \le p \le q < +\infty$, then*

$$\|a\|_q \le \|a\|_p \le k^{\frac{1}{p}-\frac{1}{q}} \|a\|_q,$$

where $\|a\|_p = \left(\sum_{i=1}^{k} |a_i|^p \right)^{1/p}$, *i.e.*

$$\left(\sum_{i=1}^{k} |a_i|^p \right)^{1/p} \le \left(\sum_{i=1}^{k} |a_i|^q \right)^{1/q} \le k^{\frac{1}{p}-\frac{1}{q}} \left(\sum_{i=1}^{k} |a_i|^p \right)^{1/p}.$$

Remark 2.6 *Note that the lower estimate in theorem 2.2 improves the estimates obtained in theorem 2.1 for all $1 \le p < \infty$, since*

$$\left[\left(\frac{1}{2} \left(\frac{2m}{n} \right)^{1/p} V_{E,p}(G) + \frac{1}{n^{1+\frac{1}{p}}} \right)^p + \frac{n-1}{n^{p+1}} \right]^{\frac{1}{p}} \ge \frac{1}{2} \left(\frac{2m}{n} \right)^{1/p} V_{E,p}(G) + \frac{1}{n^{1+\frac{1}{p}}}$$

$$\ge \frac{1}{2} \left(\frac{2m}{n} \right)^{1/p} V_{E,p}(G).$$

In addition to this the lower estimate in theorem 2.2 for all complex network G when we take $p \longrightarrow 1^+$, since we can prove that in this case

$$\left[\left(\frac{1}{2} \left(\frac{2m}{n} \right)^{1/p} V_{E,p}(G) + \frac{1}{n^{1+\frac{1}{p}}} \right)^p + \frac{n-1}{n^{p+1}} \right]^{\frac{1}{p}} \xrightarrow[p \longrightarrow 1^+]{} V_{X,1}(G).$$

References

[1] R. Albert and A.L. Barabási, *Rev. Mod. Phys.* **E74**, (2002) 47–97.

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, *Physics Reports* **424**, (2006) 175–308.

[3] S. Boccaletti, J. Buldú, R. Criado, J. Flores, V. Latora, J. Pello, M. Romance, *Chaos* ***17***, 043110 (2007).

[4] R. Criado, J. Flores, B. Hernández-Bermejo, J. Pello, M. Romance, *J. of Math. Modelling and Algorithms* **4** (2005) 307–316.

[5] R. Criado, J. Pello, M. Romance, M. Vela-Pérez, to appear in *Int. J. Bifurcat. Chaos.*

[6] M. E. J. Newman and M. Girvan, *Phys. Rev.* **E69**, 026113 (2004);

[7] W. Rudin, *Real and complex analysis* (3th. Ed.), McGraw-Hill, Boston, 1987.

[8] S. Wasserman and K. Faust, *Social Networks Analysis*, Cambridge University Press, Cambridge, 1994.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Explicit representation of Green function for 3D dimensional exterior Helmholtz equation

Pedro Cruz[1] **and Evgeny Lakshtanov**[1]

[1] *Department of Mathematics , University of Aveiro, Portugal*

emails: pedrocruz@ua.pt, lakshtanov@rambler.ru

Abstract

We have constructed a sequence of solutions of the exterior Helmholtz equation such that their restrictions form an orthonormal basis on a given surface. Coefficients of these functions depend on an explicit algebraic formulae from the coefficients of the surface. In the same way we constructed an explicit normal derivative of the Dirichlet Green function. Moreover the Dirichlet-to-Neumann operator is also constructed. We proved that normalized coefficients are uniformly bounded from zero.

Key words: explicit solution, Helmholtz exterior problem, Green function, Dirichlet-to-Neumann operator

1 Introduction

Consider $\Omega \subset \mathbb{R}^3$ with Lipschitz boundary $\partial\Omega$ and $k > 0$. The scattered field is given by Helmholtz equation and radiation condition

$$\Delta\Psi(r) + k^2\Psi(r) = 0, \quad r \in \Omega' = \mathbb{R}^3\backslash\Omega, \tag{1}$$

$$\int_{|r|=R} \left| \frac{\partial\Psi(r)}{\partial|r|} - ik\Psi(r) \right|^2 dS = o(1), \quad R \to \infty, \tag{2}$$

with Dirichlet boundary conditions,

$$\Psi(r) \equiv u_0(r), \quad r \in \partial\Omega, \quad u_0 \in C(\partial\Omega). \tag{3}$$

For example, in [1] is proved the existence and uniqueness of the solution of (1)-(3). A function $\Psi(r)$ which satisfy mentioned conditions has asymptotics

$$\Psi(r) = \frac{e^{ik|r|}}{|r|} f(q) + o\left(\frac{1}{|r|}\right), \quad r \to \infty, \quad q = r/|r| \in S^2, \tag{4}$$

where function $f(\theta,\varphi) = f(\theta,\varphi,k,u_0)$ is called *scattering amplitude* and observable $\sigma_T = \int_{S^2} |f(q)|^2 d\sigma(q)$ is called total cross section, σ is a square element of the unit sphere.

Unfortunately, analytical expressions of that observable for certain $k > 0$ exist only for few bodies of simple shapes. The ground analytical achievement by A. Ramm and S. Gutman [2] is the so called Modified Rayleigh Conjecture. In particulary, it follows that functions $Y_{lm}(\theta,\varphi)h_l(k|r|)|_{\partial\Omega}$ (spherical harmonics and spherical Hankel functions (see (5)) correspondingly) form a basis in the space $L_2(\partial\Omega, dS)$.

Let $\overline{\mathbf{N}} = \mathbf{N} \cup \{0\}$. Let $\mathcal{L}$ be the set of indexes $\mathcal{L} = \{(l,m) : l,m \in \overline{\mathbf{N}}, |m| \leq l\}$ and let $\mathbf{l} \in \mathcal{L}$ with $\mathbf{l} = (\mathbf{l}(1), \mathbf{l}(2))$. We also we use the notation $\bar{\mathbf{l}} = (\mathbf{l}(1), -\mathbf{l}(2))$ and set the order: $(l,m) > (p,r) \longleftrightarrow l > p \wedge [(l = p) \vee (|m| > |r|)] \wedge [(l = p) \vee (m > 0) \vee (m = -r)]$, so $\mathcal{L} = \{(0,0),(1,0),(1,-1),(1,1),(2,0),(2,-1),(2,1),(2,-2),(2,2),(3,0),...\}$. Let $\mathbf{o} = (0,0)$ and let operations $+$ and $-$ have the natural definition in $\mathcal{L} \times \mathbf{Z} \to \mathcal{L}$ correspondingly to the introduced order.

1.1 Surfaces with inverse radius-vector represented as finite combination of harmonics

Let $\mathcal{F}$ be a subset of functions (multi-index) $\mathcal{L}^{\overline{\mathbf{N}}}$ which have finite support and let capacity be defined by

$$|d| = \sum_{\mathbf{l}} d(\mathbf{l}), \quad \mathrm{Supp} d = \max\{\mathbf{l} : d(\mathbf{l}) \neq 0\}\,, \quad d \in \mathcal{F}.$$

Let $e_{\mathbf{l}} \in \mathcal{F}$ be defined as $e_{\mathbf{l}}(\mathbf{m}) = \delta_{\mathbf{lm}}$ (evaluated 1 only when $\mathbf{l} = \mathbf{m}$). Also set

$$C^d = \frac{|d|!}{\prod_{\mathbf{l}} d(\mathbf{l})!}, \quad I^d = \int_0^{\pi}\int_0^{2\pi} \prod_{\mathbf{l}} \left(Y_{\mathbf{l}}(\theta,\varphi)\right)^{d(\mathbf{l})} d\theta d\varphi, \quad d \in \mathcal{F}.$$

and for arbitrary function $a \in \mathcal{L}^{\mathbb{R}}$ with finite support we denote $a^d = \prod_{\mathbf{l}} a_{\mathbf{l}}^{d(\mathbf{l})}$.

Theorem 1.1 *Let a star shaped surface $\partial\Omega$ be given as a set $\{r = r(\theta,\varphi) \in \mathbb{R}^3, \theta \in [0,\pi], \varphi \in [0,2\pi)\}$ where $|r(\theta,\varphi)| = 1/\sum_{\mathbf{l} \leq (N,N)} a_{\mathbf{l}} Y_{\mathbf{l}}(\theta,\varphi)$ with $N \geq 0$ and where $\{a_{lm}\}$ are coefficients. Then*
1. Functions $\widehat{\Psi}_{\mathbf{n}}(r) = \sum_{\mathbf{k}\leq\mathbf{n}} c_{\mathbf{nk}} Y_{\mathbf{k}}(\theta,\varphi) h_{\mathbf{k}(1)}(k|r|), \quad \mathbf{n} \in \mathcal{L}$ satisfy (1),(2) and their restrictions $\{\widehat{\Psi}_{\mathbf{n}}|_{\partial\Omega}, \mathbf{n} \in \mathcal{L}\}$ form an orthonormal basis in $L_2(\partial\Omega, d\theta d\varphi)$.
Here

$$c_{\mathbf{nn}} = 1/\lambda_{\mathbf{n}}, \quad \mathbf{n} \in \overline{\mathbf{N}}$$

$$\lambda_{\mathbf{o}}^2 = g_{\mathbf{oo}} > 0, \quad \lambda_{\mathbf{n}}^2 = g_{\mathbf{nn}} - \sum_{\mathbf{k}=\mathbf{o}}^{\mathbf{n}-1} \left| \sum_{\mathbf{p}=\mathbf{o}}^{\mathbf{k}} \bar{c}_{\mathbf{kp}} g_{\mathbf{np}} \right|^2 > 0, \quad \mathbf{n} > \mathbf{o}.$$

We now define $g_{\mathbf{ij}}$, $\widehat{h}_{nm}$, $c_{\mathbf{nm}}$. Let

$$g_{\mathbf{ij}} = (-1)^{\mathbf{j}(2)} \sum_{m=0}^{\mathbf{i}(1)+\mathbf{j}(1)} \frac{1}{k^{m+2}} \left(\sum_{l=0}^{m} \widehat{h}_{\mathbf{i}(1)l} \overline{\widehat{h}}_{\mathbf{j}(1)(m-l)} \right) \sum_{d:|d|=m+2, Supp d \leq (N,N)} C^d a^d I^{d+e_{\mathbf{i}}+e_{\bar{\mathbf{j}}}}$$

where coefficients $\widehat{h}_{nj}$ are defined from the well known representation for Hankel spherical functions [3],

$$h_n(t) = \frac{e^{ikt}}{t}\frac{\sum_{j=0}^{n}\widehat{h}_{nj}t^{n-j}}{t^n} = \frac{e^{ikt}}{t}\sum_{j=0}^{n}\frac{\widehat{h}_{nj}}{t^j}, \quad \widehat{h}_{n0}=1, \quad t \neq 0, \ \textit{with} \tag{5}$$

$$\widehat{h}_{nm} = \frac{i^m}{2^m}\prod_{p=1}^{m}(n+p)\cdot\prod_{p=1}^{m}\frac{(n-m+p)}{p}, \quad 0 < m \le n;$$

$$c_{\mathbf{nm}} = \frac{1}{\lambda_{\mathbf{n}}}\left(\sum_{\mathbf{k}=\mathbf{m}}^{\mathbf{n}-1}\sum_{\mathbf{p}=\mathbf{o}}^{\mathbf{k}}\overline{c}_{\mathbf{kp}}c_{\mathbf{km}}g_{\mathbf{np}}\right), \quad \mathbf{m} < \mathbf{n}.$$

2. Consider an arbitrary function $u_0 \in L_2(\partial\Omega)$, then we have

$$\sigma_T = \frac{1}{k^2}\sum_{\mathbf{n}=\mathbf{o}}^{\infty}\sum_{\mathbf{m}\le\mathbf{n}}\left[\overline{c}_{\mathbf{nm}}\left(\sum_{\mathbf{p}\le\mathbf{n}}c_{\mathbf{np}}\overline{\widehat{u}}_{\mathbf{p}}\right)\left(c_{\mathbf{nm}}\sum_{\mathbf{p}\le\mathbf{n}}\overline{c}_{\mathbf{np}}u_{\mathbf{p}} + 2\sum_{\mathbf{m}<\mathbf{l}<\mathbf{n}}c_{\mathbf{lm}}\sum_{\mathbf{p}\le\mathbf{l}}\overline{c}_{\mathbf{lp}}u_{\mathbf{p}}\right)\right], \tag{6}$$

where

$$u_{\mathbf{p}} = \int_0^{\pi}\int_0^{2\pi}u_0(\theta,\varphi)\overline{Y}_{\mathbf{p}}(\theta,\varphi)\overline{h}_{\mathbf{p}(1)}(k|r(\theta,\varphi)|)d\theta d\varphi.$$

3. Moreover, exists numbers $C_i = C_i(k,\Omega), i = 1,2$ such that

$$c_{\mathbf{nk}} \le \frac{C_1}{\mathbf{k}(1)!}, \quad \mathbf{o} \le \mathbf{k} \le \mathbf{n}, \quad \lambda_{\mathbf{n}} > C_2, \quad \mathbf{k},\mathbf{n} \in \mathcal{L}. \tag{7}$$

4. We have weak convergence of $\frac{\partial G}{\partial n_t}$:

$$\frac{\partial G}{\partial n_t}(r,t) = \sum_{\mathbf{n}}\widehat{\Psi}_{\mathbf{n}}(r)\overline{\widehat{\Psi}}_{\mathbf{n}}(t), \quad r,t \in \Omega'.$$

Acknowledgements

This work was supported by *Centre for Research on Optimization and Control* (CEOC) from the "*Fundação para a Ciência e a Tecnologia*" (FCT), cofinanced by the European Community Fund FEDER/POCTI, and by the FCT research project PTDC/MAT/72840/2006.

References

[1] A. G. Ramm, *Scattering by Obstacles*, Dordrecht: Reidel, 1986.

[2] S. Gutman, A.G. Ramm, *Numerical implementation of the MRC method for obstacle scattering problems*, J. Phys. A: Math. Gen. **35** (2002) 8065–8074.

[3] P. M. Morse, J. Feshbach, *Method of Theoretical Physics*, McGraw-Hill, NY, 1953.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Some effective formulas for the Stirling numbers

Stefan Czerwik[1]

[1] *Institute of Mathematics, Silesian University of Technology, Gliwice, Poland*

emails: `Stefan.Czerwik@polsl.pl`

Abstract

Let $\{f_n\}$ be a sequence of real or complex numbers. Define

$$C[f_n](z) = A(z) := \sum_{n=0}^{\infty} f_n z^n. \tag{1}$$

Then we can easily verify the following

Lemma 1

$$C[n^k f_n](z) = (zD)^k A(z), \quad for\ k \in \mathbb{N}, \tag{2}$$

where $\mathbb{N}$ *denotes the set of natural numbers and* D *means the derivative.*

The proof can be done by the induction with respect to k.

Let $F\colon \mathbb{C} \to \mathbb{C}$ ($\mathbb{C}$ - the set of complex numbers be an k-times differentiable function. We define the complex numbers $t_{k,s}$ by the following way.

Definition 1 *The numbers* $t_{k,s}$ *are such that the following equality holds*

$$(zD)^k F(z) = \sum_{s=1}^{k} t_{k,s} z^s D^s F(z), \quad k \in \mathbb{N}, \tag{3}$$

and

$$t_{k,0} = 0,\ t_{k,s} = 0\ for\ s > k.$$

Now we can prove the following

Theorem 1 *We have*

$$t_{k+1,s} = s t_{k,s} + t_{k,s-1}, \tag{4}$$

for $k \in \mathbb{N}$.

Proof: We have by (3)

$$\begin{aligned}
\sum_{s=1}^{k+1} t_{k+1,s} z^s D^s F(z) &= (zD)^{k+1} F(z) \\
&= (zD)\left(\sum_{s=1}^{k} t_{k,s} z^s D^s F(z)\right) \\
&= z \sum_{s=1}^{k} t_{k,s} D(z^s D^s F(z)) \\
&= \sum_{s=1}^{k} t_{k,s} [s z^s D^s F(z) + z^{s+1} D^{s+1} F(z)] \\
&= \sum_{s=1}^{k+1} s t_{k,s} z^s D^s F(z) + \sum_{s=1}^{k+1} t_{k,s-1} z^s D^s F(z) \\
&= \sum_{s=1}^{k+1} [s t_{k,s} + t_{k,s-1}] z^s D^s F(z).
\end{aligned}$$

Hence we get the equality (4) for all $k \in \mathbb{N}$. This concludes the proof.

Remark 1 *Note that the numbers $t_{k,s}$ satisfy the same relation as the Stirling numbers of the second kind.*

Now we consider the partial difference equation

$$s_{k+1,r} = s_{k,r-1} - k s_{k,r} \tag{5}$$

defining the Stirling numbers of the first kind.

We shall find the effective formula for these numbers. We assume
$s_{k,0} = 0,\ s_{k,r} = 0\ for\ r > k.$

Take $r = 1$. Then we have

Lemma 2 *The difference equation*

$$s_{k,1} = s_{k-1,0} - (k-1) s_{k-1,1}, \quad k \in \mathbb{N} \tag{6}$$

with the condition $s_{1,1} = 1$, has exactly one solution given by the formula

$$s_{k,1} = (-1)^{k-1}(k-1)!, \quad k \in \mathbb{N}. \tag{7}$$

Having done this, one can prove the following

Lemma 3 *The difference equation*

$$s_{k,2} = s_{k-1,1} - (k-1) s_{k-1,2}, \quad k \in \mathbb{N} \tag{8}$$

with the condition $s_{2,2} = 1$, has exactly one solution given by the formula

$$s_{k,2} = (-1)^{k-2}(k-1)! \sum_{r=1}^{k-1} \frac{1}{r}, \quad k \geq 2. \tag{9}$$

Proof: Since $s_{k-1,1}$ is given by the formula (7) and $s_{2,2} = 1$, then the equation (8) has exactly one solution. One can verify that the sequence given by (9) is the solution of the equation (8). This completes the proof.

Lemma 4 *The difference equation*

$$s_{k,3} = s_{k-1,2} - (k-1)s_{k-1,3}, \quad k \in \mathbb{N} \tag{10}$$

with the condition $s_{3,3} = 1$, has exactly one solution given by the formula

$$s_{k,3} = (-1)^{k-3}(k-1)! \sum_{m=1}^{k-2} \frac{1}{m+1} \left(\sum_{r=1}^{m} \frac{1}{r} \right), \quad k \geq 3. \tag{11}$$

Proof can be done similarly to the proof of Lemma 3.

Now we are ready to state the main result of the paper.

Theorem 2 *Let $s_{k,k} = 1$ for $k \in \mathbb{N}$. Then partial difference equation (5) has exactly one solution given by the formula:*

$$s_{k,l} = (-1)^{k-l}(k-1)! \sum_{m_{l-1}=1}^{k-(l-1)} \frac{1}{m_{l-1}+l-2} \sum_{m_{l-2}=1}^{m_{l-1}} \frac{1}{m_{l-2}+l-3} \cdots \frac{1}{m_2+1} \sum_{m_1=1}^{m_2} \frac{1}{m_1}, \tag{12}$$

$k \geq l$.

Proof: Taking into account Lemmas 2, 3, 4 one can prove formula (12) by mathematical induction principle.

Remark 2 *The formula (12) is the explicit formula for the Stirling numbers of the first kind.*

Remark 3 *Some asymptotic estimates of Stirling numbers of the first kind utilizing formula (12) can be obtained.*

Remark 4 *Computer program for finding the Stirling numbers of the first kind has been established.*

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008

Modelling of the influence of diet and lifestyle on the value of bone mineral density in post-menopausal women using multivariate adaptive regression splines

F. J. de Cos Juez[1], F. Sánchez Lasheras[2], M. A. Suárez Suárez[3] and P. J. García Nieto[4]

[1] *Mining Exploitation and Prospecting Department, University of Oviedo, 33004 Oviedo, Spain*
[2] *Research Department, Tecniproject Ltd., 33004 Oviedo, Spain*
[3] *Orthopaedic Surgery Department Cabueñes Hospital, 33394 Gijon Spain*
[4] *Department of Mathematics, University of Oviedo, 33007 Oviedo, Spain*
emails: fjcos@uniovi.es, fsanchez@tecniproject.com, miguel.suarez@sespa.princast.es, lato@orion.ciencias.uniovi.es

Abstract

In this work, the application of 'multivariate adaptive regression splines' (MARS) for modelling osteoporosis is described. Osteoporosis is characterized by low 'bone mineral density' (BMD). This illness has a high-cost impact in all developed countries. The aim of this article is the development of a mathematical method able to predict the BMD of post-menopausal women taken into account only certain nutritional variables. A nutritional habits and lifestyle questionnaire is drawn up. The variables obtained from this, together with the BMD of the patients calculated by densitometry are processed using the 'principal component analysis' (PCA) algorithm in order to reduce the dimension of the database. Finally, the 'multivariate adaptive regression splines (MARS) method' is applied. It has been proved to be possible to build a MARS model in order to forecast the BMD of the post-menopausal women in function of their responses to the questionnaire. This model can be used to determine which women should take a densitometry.

Key words: Quantitative computed tomography (QCT), Body mass index (BMI), Bone mineral density (BMD), Diet history questionnaire (DHQ), Food frequency questionnaire (FFQ), Risk factor monitoring and methods branch (RFMMB),

Principal components analysis (PCA), Multivariate adaptive regression splines (MARS)
MSC 2000: 62G08, 65C60, 68T05

1. Introduction

Osteoporosis is a disease that develops without showing symptoms during its early stages. This illness is characterized by low bone mass and micro-architectural deterioration of bone tissue, which result in an increase in the risk of fractures, mainly of hips, wrists and vertebrae. The term osteoporosis and the recognition of its pathological appearance is generally attributed to the French pathologist Jean Lobstein [1] but it was the American endocrinologist Fuller Albright and collaborators who linked osteoporosis with the postmenopausal state [2]. From an economic point of view, the osteoporosis has a high cost for society, not only for the governments but also for the families. Osteoporosis is a contemporary illness that affects mainly women after menopause due to the lack of estrogens associated with it. The incidence of this illness is higher in the more developed countries, as they have populations with higher average ages.

Most authorities recommend risk-factor assessment for all postmenopausal women, followed by bone mineral density (BMD) testing for those women at highest risk for osteopenia, osteoporosis, and fractures [3-4]. There are many factors that contribute to the less than optimal identification and treatment of these patients; the difference between best practice and clinical practice with respect to the management of osteoporosis are diverse. Women at risk of osteoporotic fracture may fall into a health care gap between the obstetricians, gynaecologists, internists, and others, who are in a position to detect and treat osteoporosis, thereby preventing fractures, and the orthopaedists who are responsible for treating the fractures [5-6]. Alternatively, some women may fall victim to the failure of health care providers to initiate or alter intervention strategies, despite changes in the patient's health status that would seemingly justify such action [5].

Osteoporosis can be prevented with lifestyle advice and medication, and preventing falls in people with known or suspected osteoporosis is an established way to prevent fractures. Osteoporosis can be treated with bisphosphonates and various other medical treatments. Osteoporosis itself has no specific symptoms; its main consequence is the increased risk of bone fractures. Osteoporotic fractures are those that occur in situations where healthy people would not normally break a bone; they are therefore regarded as fragility fractures. Typical fragility fractures occur in the vertebral column, hip and wrist.

The most important risk factors for osteoporosis are age (in both men and women) and female sex; estrogens deficiency following menopause is correlated with a rapid reduction in BMD, while in men a decrease in testosterone levels has a comparable (but less pronounced) effect. While osteoporosis occurs in people from all ethnic groups, European or Asian ancestry predisposes for osteoporosis. Those with a family history of fracture or osteoporosis are at an increased risk; the heritability of the fracture as well as low bone mineral density are relatively high, ranging from 25 to 80 percent. There are at least 30 genes associated with the development of osteoporosis [7]. Those who have already had a fracture are at least twice as likely to have another fracture compared to someone of the same age and sex [8].

Modern medical facilities are equipped with monitoring, collecting and other devices which can provide inexpensive ways to collect and store data in their information systems. Huge amount of data stored in these databases need special techniques for processing, analysing, and effective use of them before these data can be helpful supports in handling medical related decision-making problems. Data mining (DM) [9-10], sometimes referred to as knowledge discovery in database (KDD) [11-12], is a systematic approach to find underlying patterns, trend, and relationships buried in data. Multivariate adaptive regression splines (MARS) [13] is one commonly used data mining technique nowadays. The aim of the present study is the creation of a predictive model of the BMD of the post-menopausal women in function of their responses to a diet and lifestyle survey. This multivariate adaptive regression splines (MARS) model could be used as a help tool for doctors in order to decide which women should take a densitometry test and which not.

2. Mathematical modelling

2.1. Multivariate adaptive regression splines (MARS)

Multivariate adaptive regression splines (MARS) is a multivariate nonparametric regression technique introduced by Friedman [13] in 1991. Its main purpose is to predict the values of a continuous dependent variable, $\vec{y}(n \times 1)$, from a set of independent explanatory variables, $\vec{X}(n \times p)$. The MARS model can be represented as:

$$\vec{y} = f(\vec{X}) + \vec{e} \tag{1}$$

where $\vec{e}$ is an error vector of dimension $(n \times 1)$.

MARS can be considered as a generalisation of 'classification and regression trees' (CART) [12], and is able to overcome some limitations of CART. MARS does not require any a priori assumptions about the underlying functional relationship between dependent and independent variables. Instead, this relation is uncovered from a set of coefficients and piecewise polynomials of degree q (basis functions) that are entirely "driven" from the regression data $(\vec{X}, \vec{y})$. The MARS regression model is constructed by fitting basis functions to distinct intervals of the independent variables. Generally, piecewise polynomials, also called splines, have pieces smoothly connected together. In MARS terminology, the joining points of the polynomials are called knots, nodes or breakdown points. These will be denoted by the small letter *t*. For a spline of degree *q* each segment is a polynomial function. MARS uses two-sided truncated power functions as spline basis functions, described by the following equations [14]:

$$\left[-(x-t)\right]_+^q = \begin{cases} (t-x)^q & if\ x<t \\ 0 & otherwise \end{cases} \tag{2}$$

$$\left[+(x-t)\right]_+^q = \begin{cases} (t-x)^q & if\ x\geq t \\ 0 & otherwise \end{cases} \tag{3}$$

where $q(\geq 0)$ is the power to which the splines are raised and which determines the degree of smoothness of the resultant function estimate.

The MARS model of a dependent variable $\vec{y}$ with M basis functions (terms) can be written as [14-15]:

$$\hat{\vec{y}} = \hat{f}_M(\vec{x}) = c_0 + \sum_{m=1}^{M} c_m B_m(\vec{x}) \tag{4}$$

where $\hat{\vec{y}}$ is the dependent variable predicted by the MARS model, c_0 is a constant, $B_m(\vec{x})$ is the *m*-th basis function, which may be a single spline basis functions, and c_m is the coefficient of the *m*-th basis functions.

Both the variables to be introduced into the model and the knot positions for each individual variable have to be optimized. For a data set $\vec{X}$ containing *n* objects and *p* explanatory variables, there are $N = n \times p$ pairs of spline basis functions, given by Eqs. (2) and (3), with knot locations x_{ij} ($i = 1,2,...,n;\ j = 1,2,...,p$).

A two-step procedure is followed to construct the final model. First, in order to select the consecutive pairs of basis functions of the model, a two-at-a-time forward stepwise procedure is implemented [13-15]. This forward stepwise selection of basis function leads to a very complex and overfitted model. Such a model, although it fits the data well, has poor predictive abilities for new objects. To improve the prediction, the redundant basis functions are removed one at a time using a backward stepwise procedure. To determine which basis functions should be included in the model, MARS utilizes the generalized cross-validation [12-15] (GVC), The GVC is the mean squared residual error divided by a penalty dependent on the model complexity. The GVC criterion is defined in the following way:

$$GVC(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_M(\vec{x}_i)\right)^2}{\left(1 - C(M)/n\right)^2} \tag{5}$$

where $C(M)$ is a complexity penalty that increases with the number of basis functions in the model and which is defined as:

$$C(M) = (M+1) + d\,M \qquad (6)$$

Where M is the number of basis functions in Eq. (4), and the parameter d is a penalty for each basis function included into the model. It can be also regarded as a smoothing parameter. Large values of d lead to fewer basis functions and therefore smoother function estimates. For more details about the selection of the d parameter, see Ref. [13]. In our studies, the parameter d equals 2, and the maximum interaction level of the spline basis functions is restricted to 3.

The main steps of the MARS algorithm as applied here can be summarized as follows [14]:

1. Select the maximal allowed complexity of the model and define the d parameter.
   - ➢ Forward stepwise selection:
2. Start with the simplest model, i.e. with the constant coefficient only.
3. Explore the space of the basis functions for each explanatory variable.
4. Determine the pair of basis functions that minimizes the prediction error and include them into the model.
5. Go to step 2 until a model with predetermined complexity is derived.
   - ➢ Backward stepwise selection:
6. Search the entire set of basis functions (excluding the constant) and delete from the model the one that contributes least to the overall goodness of fit using the GCV criterion.
7. Repeat 5 until GCV reaches its maximum.

The predetermined complexity of MARS model in step 3 should be considerably larger than the optimal (minimal GCV) model size M^*, so choosing the predetermined complexity of the model as more than $2M^*$ is enough in general [13].

2.2 Prediction ability of the MARS model

The prediction ability of the MARS model can be evaluated in terms of the ‘root mean squared error of cross-validation’ (RMSECV) and the squared leave-one-out correlation coefficient (q^2). To compute RMSECV, one object is left out from the data set and the model is constructed for the remaining $n-1$ objects. Then the model is used to predict the value for the object left out. When all objects have been left out once, RMSECV is given by [15]:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{-i}\right)^2}{n}} \qquad (9)$$

where y_i is the value of dependent variable of the *i*-th object, $\hat{y}_{-i}$ is the predicted value of the dependent variable of the *i*-th object with the model built without the *i*-th object.

The value of q^2 is given as:

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{10}$$

where $\bar{y}$ is the mean value of the dependent variable for all *n* objects.

2.3 The importance of the variables in the MARS model

Once the MARS model is constructed, it is possible to evaluate the importance of the explanatory variables used to construct the basis functions. Since each explanatory variable can be incorporated into different basis functions, the importance of the variable is expressed as its contribution to the goodness of fit of the model. The scoring of the importance of variables in the MARS model is similar to the leave-one-out cross-validation concept. To calculate variable importance scores, MARS refits the model after deleting all terms involving the variable at issue and calculating the reduction in goodness of fit. The importance of the variables is a relative measure and scaled between 0 and 1. The most important variable is the one that, when dropped, decreases the model fit the most and it receives the highest score, i.e. 1. The less important variables receive the lower scores, which is the ratio of the reduction in goodness of fit of these variables to that of the most important variable.

3. Empirical study

3.1 Population

The women in this study were members of a random healthy population. The examinations were performed between the month of September 2004 and June 2007. The questionnaire was offered to post-menopausal women between 50 and 69 years old. All the patients gave their written approval to take part in the present study. The research methodology was approved by the ethical committee of the Medical Institution.

Those women that presented one or more of the following criteria were excluded from the study: chronic renal insufficiency, any kind of endocrinopathy, bone metabolic illness, chronic hepatopathy, any kind of neoplasia and treatment with glucocorticosteroid.

Table 1 lists the basic characteristics of the population of the study, including parameters such as height, weight, Body Mass Index (BMI), age at which they reached menopause, etc. The BMD of the women that take part in the study was measured with a dual energy densitometer (DXA).

Table 1. Basic characteristics of patients (size of the sample = 305).

Variable	Average	Standard deviation (σ)	Mininum value	Maximum value
Height (m)	1.56	0.1127	1.31	1.78
Weight (kg)	69.32	7.83	46	95
BMI	28.92	5.24	18.19	41.95
Age	57.91	5.34	50	69
Age at menopause (years)	52.82	43.31	35	62
Number of children	25.47	14.55	0	6
Number of abortion	0.28	0.53	0	2
Number of pregnancies	28.23	15.04	0	7

3.2 Nutritional habits and life style questionnaire

In order to find a relationship between the nutritional habits and the lifestyle of the post-menopausal women, a specific questionnaire was designed. The nutritional questions of this questionnaire were taken from the Diet History Questionnaire (DHQ) [16]. This is a Food Frequency Questionnaire (FFQ) developed by staff at the Risk Factor Monitoring and Methods Branch (RFMMB). The questionnaire contains some demographic questions about the patient, and others related to their health and lifestyle. The main block of questions in the survey is about dietary habits. The survey was distributed by hand to all the patients of the sample.

The completion of the questionnaire by the patients, gave to the researchers a list of 39 variables including the BMD value for each of the patients. All the variables considered, according to previous medical researches have influence in the evolution of the osteoporosis disease [17]. These variables include factors such as calcium intake, proteins intake, number of pregnancies, height, body mass index (BMI), etc.

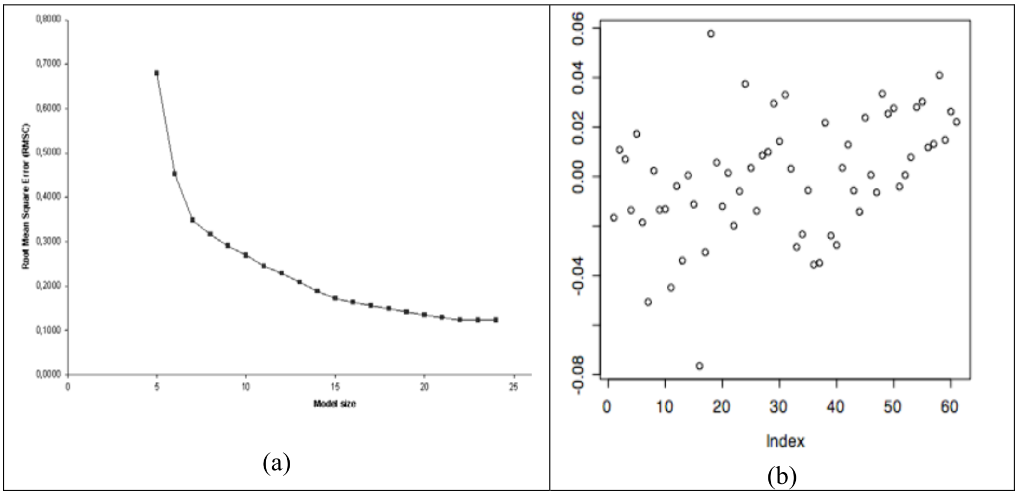
4. Results and discussion

In order to detect the influence of the 38 variables over the BMD and discard the less relevant, the Principal Components Analysis Algorithm (PCA) was employed [10-12, 18]. The PCA algorithm not only allows variables with a high degree of dependence between each other to be discarded but also performs a first evaluation of the relative influence that the variables would have on the parameter to model: the BMD, in this case.

The definitive list of variables taken into consideration after the application of the PCA algorithm is listed in Table 2. The total number of prediction variables used to build the MARS model is 12. In this work we use a second-order MARS, so that the basis functions of the models consist of linear and second-order splines and the maximum number of terms was not limited (no pruning).

Table 2. Reduced set of variables used in the study and the value of their 5 first principal components.

Name of the variable	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
BMI	0.108	0.046	0.012	0.057	0.114
CALCIUM	0.557	0.237	0.228	-0.207	0.138
CHOLESTEROL	0.635	0.321	-0.385	-0.156	0.173
CARBOHIDRATES	0.381	0.695	0.845	0.363	0.684
ENERGY	0.281	-0.396	-0.144	0.023	0.012
FATS	0.364	0.147	0.366	0.162	-0.203
FOLATE	0.423	0.004	-0.008	0.010	0.000
PHYS_ACT	-0.572	0.136	0.185	0.125	-0.018
PREGNANCIES	0.358	0.351	0.029	0.210	0.090
PROTEINS	0.760	0.340	0.576	0.963	0.062
SUN_EXPOSURE	0.643	0.093	0.597	0.485	0.266
VITAMIN_D	0.560	0.840	0.038	0.079	0.342



(a)

(b)

Figure 1: (a) Root mean square error (RMSC) versus model complexity and (b) residuals of the empirical data versus the MARS model results.

Table 3. List of basis functions B_i of the MARS model and their coefficients a_i .

B_i	Definition	a_i
B_1	1	0.73830
B_2	$(\text{BMI} - 26.95492)$	0.02621
B_3	$(\text{CALCIUM} - 585)$	0.01019
B_4	$(\text{CHOLESTEROL} + 0.45257)$	0.00001
B_5	$(\text{CARBOHIDRATES} + 13.25585)$	0.00235
B_6	$(\text{ENERGY} - 800)$	-0.02240
B_7	$(\text{FATS} + 12.82473)$	0.00464
B_8	$(\text{FOLATE} - 250.14585)$	0.00356
B_9	$(\text{PHYS\_ACT} - 1)$	0.05167
B_{10}	$(\text{PREGNANCIES} - 2)$	0.02203
B_{11}	$(\text{PROTEINS} - 69)$	0.01091
B_{12}	$(\text{SUN\_EXPOSURE} - 1)$	0.03205
B_{13}	$(\text{VITAMIN\_D} - 69)$	-0.00232
B_{14}	$(\text{BMI} - 26.95492) \cdot (\text{CARBOHIDRATES} + 13.25585)$	0.00299
B_{15}	$(\text{BMI} - 26.95492) \cdot (\text{PREGNANCIES} - 2)$	0.00147
B_{16}	$(\text{BMI} - 26.95492) \cdot (\text{VITAMIN\_D} - 69)$	0.00006
B_{17}	$(\text{CALCIUM} - 585) \cdot (\text{CARBOHIDRATES} + 13.25585)$	0.00010
B_{18}	$(\text{CALCIUM} - 585) \cdot (\text{FOLATE - 250.14585})$	0.00142
B_{19}	$(\text{CALCIUM} - 585) \cdot (\text{PREGNANCIES - 2})$	-0.01919
B_{20}	$(\text{CALCIUM} - 585) \cdot (\text{VITAMIN\_D- 69})$	0.00865
B_{21}	$(\text{CARBOHIDRATES} + 13.25585) \cdot (\text{ENERGY - 800})$	0.00116
B_{22}	$(\text{CARBOHIDRATES} + 13.25585) \cdot (\text{PREGNANCIES - 2})$	0.00399
B_{23}	$(ENERGY - 800) \cdot (\text{FATS} + 12.82473)$	0.00942
B_{24}	$(ENERGY - 800) \cdot (\text{PREGNANCIES} - 2)$	0.00007

The MARS model is computed using an 80% ($n = 244$) of the data selected by random. The validation of the model obtained has been performed with the other 20% of data ($n = 61$) by means of the residuals calculating as the difference between empirical data and the obtained data through the MARS model. Figure 1 (a) shows the root mean square error (RMSC) of the model as a function of the model complexity (number of terms). It can be observed that since term 20, the addition of more does not reduce significantly the RMSC. Figure 1 (b) shows the residuals of the empirical data (the other 20%) versus the MARS model results. Table 3 shows a list of the 24 main basis functions of the MARS models and their coefficients a_i. In spite of the model was computed without pruning, only the first 24 terms are shown in Table 3 because as it has being stated before the reduction in the RMSC with the addition of more terms is not significant.

According to what is shown in Table 3, the most important variables for the prediction of osteoporosis in post-menopause women are as follows: BMI, PREGNANCIES, PROTEINS, CALCIUM, ENERGY, PHYS_ACT, SUN EXPOSURE. From a mathematical point of view, its importance is defined by the value of their coefficients.

Finally, it can be remarked that the model obtained was applied to a new sample of patients ($n = 20$) showing a sensitivity of 90% and specificity of 85%.

5. Conclusions

MARS exhibits the capability of modelling complex relationship among variables without strong model assumptions. Besides, MARS does not need long training process and hence can save lots of modelling time when the data set is huge. Finally, one strong advantage of MARS over other classification techniques is that the resulting model can be easily interpreted as in our case.

Due to the nonlinear and local character of the MARS model, we have being able to model very complex relationships in the data set. Although the interpretation of the basis functions is difficult, it is possible to evaluate the importance of certain variables for the model and thus to understand which variables have a high influence over the osteoporosis. The variables that have been found as important in the MARS model as well as the interactions they provide are in line with previous researches.

Acknowledgements

The authors express deep gratitude to Department of Natural Resources and Environmental Engineering at Vigo University and to Departments of Mining Exploitation and Prospecting and Mathematics at Oviedo University as well as to Tecniproject Ltd Company for its computational support.

References

[1] J.G.C.F.M. LOBSTEIN, *Lehrbuch der pathologischen Anatomie*, Bd II, Stuttgart, 1835.

[2] F. ALBRIGHT, E. BLOOMBERG, P.H. SMITH, *Postmenopausal osteoporosis*, Trans. Assoc. Am. Physicians **55** (1940) 298-305.

[3] NORTH AMERICAN MENOPAUSE SOCIETY, *Management of postmenopausal osteoporosis: position statement of the North American Menopause Society,* Menopause **9** (2002) 84-101.

[4] J.P. BROWN, R.G. JOSSE, *Clinical practice guidelines for all the diagnosis and management of osteoporosis in Canada,* Scientific Advisory Council of the Osteoporosis Society of Canada, CMAJ **167** (2002) Suppl. 1-34.

[5] S.E. ANDRADE, S.R. MAJUMDAR, K.A. CHAN, D.S. BUIST, A.S. GO, M. GOODMAN, *Low frequency of treatment of osteoporosis among postmenopausal women following a fracture*, Arch. Intern. Med. **163** (2003) 2052-2057.

[6] C. SIMONELLI, K. KILLEEN, S. MEHLE, L. SWANSON, *Barriers to osteoporosis identification and treatment among primary care physicians and orthopedic surgeons.* Mayo Clin. Proc. **77** (2002) 334-338.

[7] H.W. CONSUELO, J.B. STANLEY, *Prevention of osteoporotic fractures in the elderly*, Am. J. Med. **118** (2005) 1190-1195.

[8] D.D. PIERRE, *Treatment of postmenopausal osteoporosis*, The Lancet **359** (2002) 2018-2026.

[9] P.-N. TAN, M. STEINBACH, V. KUMAR, *Introduction to Data Mining*, Addison Wesley, New York, 2005.

[10]I.H. WITTEN, E. FRANK, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, New York, 2005.

[11]J. HAN, M. KAMBER, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, New York, 2005.

[12]T. HASTIE, R. TIBSHIRANI, J.H. FRIEDMAN, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2003.

[13]J.H. FRIEDMAN, *Multivariate adaptive regression splines*, Ann. Stat. **19** (1991) 1-141.

[14]S. SEKULIC, B.R. KOWALSKI, *MARS: a tutorial*, J. Chemom. **6** (1992) 199-216.

[15]J.H. FRIEDMAN, C.B. ROOSEN, *An introduction to multivariate adaptive regression splines*, Statistical Methods in Medical Research **4** (1995) 197-217.

[16]NATIONAL INSTITUTES OF HEALTH, *Diet History Questionnaire*, National Cancer Institute, Applied Research Program, 2002.

[17]R. MARCUS, D. FELDMAN, D. NELSON, C.J. ROSEN, *Osteoporosis*, Academic Press, New York, 2007.

[18]P.W. HOWE, *Principal components analysis of protein structure ensembles calculated using NMR data*, J. Biomol. NMR. **20** (2001) 61-70.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008

Dynamic Case Based Planning

Juan F. De Paz, Sara Rodríguez, Juan M. Corchado, Javier Bajo

Departamento de Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, España
Email: {fcofds, srg, corchado, jbajope}@usal.es

Abstract

This paper presents a CBP-BDI planning model which incorporates a novel artificial neural network. The CBP-BDI model, which is integrated within an agent, is the core of a Multiagent System that allows managing the security in industrial environments. The proposed model uses Self-Organized Maps to calculate optimum routes for the security guards. Besides, some technologies of Ambient Intelligence such as RFID and Wi-Fi are used to develop the intelligent environment that has been tested and analyzed in this paper.

Key words: Multiagent Systems, Case-Based Reasoning, Cased-Based panning, BDI, Ambient Inteligent, Selft-Organized Maps, RFID

1 Introduction

During the last decades, there has been an important evolution in the management of business using Artificial Intelligence techniques. But there are some aspects that still need to be improved, especially in techniques and technology for monitoring the workers activities in more efficient ways. It is necessary to establish security policies to manage risks and control hazardous events, providing better working conditions and an increase in productivity. Implementation of time control systems has a good influence in productivity, since the workers optimize their potential and enhance the process where they collaborate. The remote monitoring is becoming increasingly common in industrial scenarios, where recent studies [1] reveal that at least 3% of working shifts time is spent because of lack of time control system, allowing supervisors to observe the behaviour of remote workers and the state of facilities.

Multi-agent systems have been recently explored as supervision systems, with the flexibility to be implemented in a wide diversity of devices and scenarios, including industrial environments. This has prompted the use of ubiquitous computing [10], which constitutes the most optimistic approach to solve the challenge to create strategies that allow the anticipation and prevention of problems on automated environments [2]. The agents have several capabilities such us autonomy, learning, reasoning. They allow developing applications in dynamic and flexible environments. These capabilities can be modelled in different ways and with different tools [8], with the use of Case Based Reasoning (CBR) systems as a possibility.

This paper focuses on presenting a hybrid CBR-based (Case Based Reasoning) deliberative agent architecture BDI (Beliefs, Desires, Intentions) [13] that incorporates a specialized planning mechanism Case-Based Planning CBP [5] to implement the retrieve, reuse, revise and retain stages of the CBR system. These hybrids architecture will be called CBP-BDI [7] [5]. BDI agents use mental aptitudes as beliefs, desires and intentions to develop intentional processes. A CBR system uses past experiences to resolve new problems and executes a sequential cycle composed of four stages. The integration of CBR systems within BDI agents provides a powerful tool to resolve problems in dynamic environments.

The use of wireless technologies, such as GPRS (General Packet Radio Service), UMTS (Universal Mobile Telecommunications System), RFID (Radio-frequency identification) [3], Bluetooth, etc., make possible to find better ways to provide mobile services and also give the agents the ability to communicate using portable devices (e.g. PDA's and cellular phones) [4].

In section 2 the basics of CBR systems are presented. Next, in section 3, the CBP-BDI model proposed in this work is explained in detail. Then, in section 4, a case study is presented, describing the main technologies used to schedule and monitor security guards surveillance routes on industrial environments and finally, in section 5 results and conclusions are exposed.

2 CBR and CBP

Case-based Reasoning (CBR) is a type of reasoning based on the use of past experiences [9] to resolve new problems. CBR systems solve new problems by adapting solutions that have been used to solve similar problems in the past, and learn from each new experience. The primary concept when working with CBR's is the concept of case. A case can be defined as a past experience, and is composed of three elements: A problem description, which describes the initial problem; a solution, which provides the sequence of actions carried out in order to solve the problem; and the final state, which describes the state achieved once the solution was applied. A CBR manages cases (past experiences) to solve new problems. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain.

Case-based planning (CBP) is a variation of CBR which consists of the idea of planning as remembering [5]. In CBP, the solution proposed to solve a given problem is a plan, so this solution is generated taking into account the plans applied to solve similar problems in the past. The problems and their corresponding plans are stored in a plans memory. In practice, what is stored is not only a specific problem with a specific solution, but also additional information about how the plans have been derived. The formal description of a case-based planner can be formalized as a 3-tuple *<I,G,Op>*:

- *I* is a set of formulae describing the initial state.
- *G* is a set of formulae describing the goal specification.
- *Op* is the set of operators (also called actions) that can be applied in a plan. Every action *a Є Op* is described in terms of pre-conditions *Ca* (what has to be fulfilled in order to the action can be executed) and port-conditions *Ea* (what has to be fulfilled after the execution of the action).

A plan *P* is a tuple *<S,B,O,L>*:

- *S* is the set of plan actions. There are two special actions: t_I, those whose effects are *I*, that is, the initial state; and t_G, those actions whose pre-conditions are *G*, that is, the goal specification.
- is an ordering relation on *S* allowing to establish an order between the plan actions. t_I is always the first action and t_G is the last action. If the ordering relation is total, P is a linear plan, whereas if it is a partial-order relation, *P* is a non-linear plan.
- *B* is a set that allows describing the bindings and forbidden bindings on the variables appearing in *P*.
- *L* is a set of casual links of the form $s \xrightarrow{p} s'$, where *s, s' Є S*, *p Є Es* and *p Є Cs'*. That is, relations allowing to establish a link between plan actions.

A plan *P* constitutes the solution generated to solve a planning problem when for each action $s \neq t_I$, for each *p Є Cs* there exists a causal link $s \xrightarrow{p} s'$ and for each action $s \neq t_G$ there exists at least a causal link $s \xrightarrow{q} s''$. In the case that the planner is interested in retaining the failures or unexpected situations during the plan, these failures or situations are represented as a set of formulae *F*.

3 CBP-BDI

Agents with BDI architecture have their origins in the *practical reasoning* of the traditional philosophy. These agents are supposed to be able to decide in each moment what action to execute according to their objectives. The practical reasoning undergoes two phases: in the first one the goals are defined and in the second one it is defined how to achieve such goals A representation based on an action requires an agent architecture in which the way to acquire and process the knowledge of the world, at the reasoning stage, is closely related to the way in which plans are constructed and used, in the phase of execution. In this section, we show how such a requirement can be achieved through a BDI agent model [7][13]. The terminology used is the following.

- The environment M and the changes that are produced within it, are represented from the point of view of the agent. Therefore, the world can be defined as a set of variables that influence a problem faced by the agent

$$M = \{\tau_1, \tau_2, \cdots, \tau_s\} \text{ with } s < \infty \qquad (1)$$

- The beliefs are vectors of some (or all) of the attributes of the world taking a set of concrete values

$$B = \{b_i / b_i = \{\tau_1^i, \tau_2^i, \cdots, \tau_n^i\}, n \leq s \quad \forall i \in N\}_{i \in N} \subseteq M \qquad (2)$$

- A state of the world $e_j \in E$ is represented for the agent by a set of beliefs that are true at a specific moment in time t.

  Let $E = \{e_j\}_{j \in N}$ set of status of the World if we fix the value of t then

$$e_j^t = \{b_1^{jt}, b_2^{jt}, \cdots b_r^{jt}\}_{r \in N} \subseteq B \quad \forall j, t \qquad (3)$$

- The desires are imposed at the beginning and are applications between a state of the current world and another that it is trying to reach

$$\begin{array}{cccc} d: & E & \rightarrow & E \\ & e_0 & \rightarrow & e^* \end{array} \qquad (4)$$

- Intentions are the way that the agent's knowledge is used in order to reach its objectives. A desire is attainable if the application i, defined through n beliefs exists:

$$\begin{array}{cccc} i: & \overset{n)}{BxBx\cdots xBxE} & \rightarrow & E \\ & (b_1, b_2, \cdots\cdots\cdots, b_n, e_0) & \rightarrow & e^* \end{array} \qquad (5)$$

  In our model, intentions guarantee that there is enough knowledge in the beliefs base for a desire to be reached via a plan of action.

- We define an agent action as the mechanism that provokes changes in the world making it change the state,

$$\begin{array}{cccc} a_j: & E & \rightarrow & E \\ & e_i & \rightarrow & a_j(e_i) = e_j \end{array} \qquad (6)$$

- Agent plan is the name we give to a sequence of actions that, from a current state e_0, defines the path of states through which the agent passes in order to reach the other world state.

$$\begin{array}{cccc} p_n: & E & \rightarrow & E \\ & e_0 & \rightarrow & p_n(e_0) = e_n \end{array}$$

$$p_n(e_0) = e_n = a_n(e_{n-1}) = \cdots = (a_n \circ \cdots \circ a_1)(e_0) \quad p_n \equiv a_n \circ \cdots \circ a_1 \qquad (7)$$

Below, the attributes that characterise the plans in the case base are presented, which allow us to relate BDI language with the interest parameters within a CBP. A constraint satisfaction problem (CSP) planning problem is considered in order to lend the model generality. These kinds of problems do not only search for solutions but also have to conform to a series of imposed restrictions. Based on the theory of action, the set of objectives for a plan and the resources available are selected as a variable upon which the CSP problems impose the restrictions. A plan *p* is expressed as *p=<E, O, O', R, R'>,* where:

E is the environment, but it also represents the type of problem faced by the agent, characterised by $E = \{e_0, e^*\}$, where e_0 represents the starting point for the agent when it begins a plan, and $e*$ is the state or states that it is trying to attain.
O indicates the objectives of the agent and *O'* are the results achieved by the plan. *R* are the total resources and *R'* are the resources consumed by the agent. Table 1 shows the indicators derived from the attributes described above, used to identify and contrast the quality of the different plans (# means cardinal of a set).

Indicator	**Formulae**
Efficacy of the plan: relationship between objectives attained and objectives proposed	$E_f = \frac{\#(O' \cap O)}{\#O}$
Cost of the plan: relationship between the resources used and the resources available	$C = \frac{\#R'}{\#R}$
Efficiency of the plan: relationship between the objectives attained and the resources consumed	$E_{ff} = \frac{\#(O' \cap O)}{\#R'}$

Table 1. Indicators of plan quality.

If a problem $E= \{e_0, e_1\}$ has been defined, a plan *p* to solve the problem can be characterised by the relationships between the objectives reached and the resources consumed between both states. The general functioning process is derived by following the typical phases of a case based system [14](eliminating the revision phase, since it can be external to the system needing the intervention of an expert). The reasoning process of this kind of system carries out the following four sequential stages (noticing that the revision stage has been eliminated: it usually is carried out by an expert and is external to the system):

- **Retrieval:** Given a state of the perceived world e_0 and the desire that the agent encounters in a state $e_0 \neq e^*$, the system searches in the case base for plans that have resolved similar problems in the past.
- **Adaptation/Reuse:** From the previous phase, a set of possible solutions for the agent $\{p_1, \ldots, p_n\}$ is obtained. In this phase, in accordance with the planning model *G*, the system uses the possible solutions to propose a solution p^* (8).

$$G(e_0, p_1, \cdots, p_n) = p^* \qquad (8)$$

- **Learning/Retain:** The plan proposed may achieve its objective or fail in its development. The information on the quality of the final plan in the $w_f(p^*)$ cycle is stored for the future and is directly proportional (i) to the initial value of $w_i(p^*)$, and (ii) to the "rate of use" $\alpha(N)$, where N is the number of times that the plan has been used in the past.

$$w_f(p^*) = w_i(p^*)\alpha(N) \qquad (9)$$

The model proposed conforms to the conditions required in order to obtain a representation and reasoning based on the action [16]. The capabilities of the hybrid system restrict what kind of plans can be generated. Plans structure and world representation can be easily adapted to a wide range of problems.

4 Case Study

A multi-agent system has been developed to provide control over the activities performed by the staff responsible for overseeing the industrial environments. The agents in the system calculate the surveillance routes for the security guards depending on the working shifts, the distance to be covered in the facilities and the security guards available. Considering this latter feature, the system has the ability to re-plan the routes automatically. A supervisor can set the possible routes, defining the areas that must be supervised. It is also possible to track the workers activities (routes completion) over the Internet.

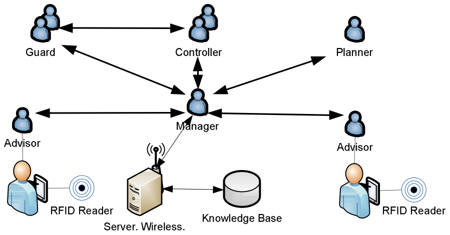


Figure 1. System Structure

Radiofrequency Identification (RFID) is a key technology in this development. It can be used to electronically identify, track, and store information about products, items, components or people. Once defined the system structure, shown on Figure 1, it is possible to define the five different kinds of agents:

- Guard Agent. It is associated to each PDA. Manages the portable RFID readers to get the RFID tags information on every control point. Communicates with Controller Agents to check the accomplishment of the assigned surveillance routes, to obtain new routes, and also to send the RFID tags information via Wi-Fi.
- Manager Agent. Controls the rest of agents in the system. Manages the connection and disconnection of Guard Agents to determine the available security guards available. The information is sent to the Planner Agent to generate new surveillance routes. It also receives incidences (omitted control points, route changes, new security guard connected/disconnected, security guards notifications, etc.) from the Controller Agents and Guard Agents and, depending on its priority, informs the Advisor Agent. Manager Agent stores all the system information (incidences, time, data, control points, route status, etc.) into a database.
- Planner Agent. Generates automatically the surveillance routes which are sent to the Manager Agent to distribute them among the security guards.
- Controller Agent. Monitors the security guards activities by means of the control points checked. Once a surveillance route is generated by the Planner Agent, the average time to reach each control point is calculated.

The Controller Agent also handles the associated route incidences and sends them to the Manager Agent.

- Advisor Agent. Administer the communication with the supervisors (person). Receive from the Manager Agent the incidences, and decide if are sent to the supervisor. Incidences can be sent via Wi-Fi, SMS or GPRS.

The agents of the system react to the events in the environment. The most important agent in the system is the Planner agent, which incorporates the CBP-BDI model. Table 2 shows the structure for a plan will considered by the Planner.

Task Field	Field Type
taskList	ArrayList of Position
numberAgents	Time

Table 2. Task structure

The information stored for each route is shown in Table 3.

Task Field	Field Type
taskList{Position, Estimated Arrive, Arrive}	ArrayList of Task

Table 3. Route

The variation of the agent plan $p_A(t)$ will be provoked essentially by: the changes that occur in the environment and that force the initial plan to be modified, and the knowledge from the success and failure of the plans that were used in the past, and which are favoured or punished via learning. The planning is carried out through a neural network based on the Kohonen network [12]. Each of the phases of the CBP-BDI planner are explained in detail in the following sub-sections:

4.1 Retrive

In this phase the most similar plan resolved in the past including all the control points indicated in the new problem is recovered. The information of the plan is given for the following record.

$$< T = \{x_i / x_i = (x_{i1}, x_{i2}), i = 1...n\}, g > \qquad (10)$$

Being x_i the control point i that it will be visited, (x_{i1},x_{i2}) the coordinates of point I and g the number of security guards. The routes r_i recovered follow the equation.

$$R = \{r_i\}\ i = 1...g \text{ where } r_i \subseteq T, r_i \cap r_j = \phi \forall i \neq j\ j = 1...g \qquad (11)$$

4.2 Reuse

If $R = \{\phi\}$ or the user establishes that he wishes to make a new distribution of the routes, the system will create a new allocation for the control points among routes. The following algorithm allows distributing the points. For surveillance routes calculation, the system takes into account the time and the minimum distance to be covered. So it is necessary a proper control points grouping and order on each group. The planning mechanism uses Kohonen SOM (Self

Organizing Maps) neural networks with the k-means learning algorithm [11] to calculate the optimal routes and assign them to the available security guards. Neural networks allow the calculus of variable size data collections, and reduce the time and distances to be covered. The distribution of the control points must follow the equation. In addition, the control points can be changed on each calculation, so the surveillance routes are dynamic, avoiding repetitive patterns.

Once the distribution of the points among routes r_i has been made, the CBP-BDI starts spreading the control points among the available security guards. Then, the optimal route for each one is calculated using a modified SOM neural network. The modification is done through a FYDPS neural network, changing the neighbourhood function defined in the learning stage of the Kohonen network. The new network has two layers: IN and OUT. The IN layer has two neurons, corresponding the physical control points coordinates. The OUT layer has the same number of control points on each route [10]. Be $x_i \equiv (x_{i1}, x_{i2}) \quad i = 1, \cdots . N$ the i control point coordinates and $n_i \equiv (n_{i1}, n_{i2}) \quad i = 1, \cdots, N$ the i neuron coordinates on $\Re^2$, being N the number of control points in the route. So, there are two neurons for the IN layer and N neurons for the OUT layer. The weight actualization formula is defined by the following equation:

$$w_{ki}(t+1) = w_{ki}(t) + \eta(t) g(k,h,t)(x_i(t) - w_{ki}(t)) \qquad (12)$$

Be w_{ki} the weight that connect the IN layer i neuron with the OUT layer k neuron. t represents the interaction; $\eta(t)$ the learning rate; and finally, $g(k,h,t)$ the neighbourhood function, which depends on three parameters: the winner neuron, the actual neuron, and the interaction.

A decreasing neighbourhood function is considered with the number of interactions and the winner neuron distance.

$$g(k,h,t) = Exp\left[\left(-\frac{|k-h|}{N/2} \right)^{\frac{\underset{\substack{i,j\in\{1,\cdots,N\} \\ i\neq j}}{Máx}\{f_{ij}\} - \sqrt{(n_{k1}-n_{h1})^2 + (n_{k2}-n_{h2})^2}}{\underset{\substack{i,j \\ i\neq j}}{Máx}\{f_{ij}\}}} - \lambda \frac{|k-h| t}{\beta N} \right] \qquad (13)$$

λ and β are determined empirically. The value of λ is set to 1 by default, and the values of β are set between 5 y 50. t is the current interaction. Its value is obtained by means of β N. $Exp[x] = e^x$, where N is the number of control points. f_{ij} is the distance between two points i and j. Finally, $Max\{f_{ij}\}$ represents the maximum distance that joins those two points.

To train the neural network, the control points groups are passed to the IN layer, so the neurons weights are similar to the control points coordinates. When all the process concludes, there is only one neuron associated to each control point. To determine the optimal route, the i neuron is associated with the i+1 neuron, from i=1, 2, …, N, covering all the neurons vector. A last interval is added to complete the route, associating the N neuron with the i neuron. The learning rate depends on the number of interactions, as can be seen on the following equation:

$$\eta(t) = Exp\left[-\sqrt[4]{\frac{t}{\beta N}}\right] \qquad (14)$$

The neurons activation function is the identity. When the learning stage ends, the winner neuron for each point is determined, so each point has only one neuron associated. The optimal route is then calculated following the weights vector. This vector is actually a ring, where the n_1 neuron is the next n_N neuron. Initially considering a high neighbourhood radius, the weights modifications affect the nearest neurons. Reducing the neighbourhood radius, the number of neurons affected decrease, until just the winner neuron is affected.

The initial number of interactions is $T_1 = \beta N$ in the first stage. When $t = \beta N$, the weights of the possible couple of neurons are changed from the neurons ring obtained. If the distance is optimized, the number of interactions is reduced to continue the learning. In the Z phase, the total number of interactions is:

$$T_Z = T_{Z-1} - \frac{T_{Z-1}}{Z} \qquad (15)$$

The objective of these phases is to avoid the crossings. Once all interactions are concluded, the distance obtained is analyzed to determine if it is the optimal distance. So, the recoil in the number of interactions is reduced each time, obtaining a maximum number of interactions, although the value is variable. Figure 2 shows the routes calculated for one and two security guards.

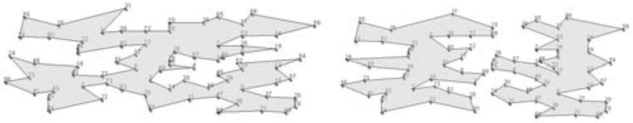


Figure 2. Planned routes for one (left) and two (right) security guards

4.3 Retain and Revise

In this case study, the final routes for the users were stored when they were successfully completed.

5 Results and Conclusions

The system presented on this paper has been implemented and tested over experimental and controlled scenarios. Simulations have been done to calculate surveillance routes and monitor the accomplishment of each one. The results obtained have shown that it is possible to find out the necessary number of security guards depending on the surveillance routes calculated by the system.

To evaluate the system efficiency, a comparison after and before the prototype implementation was done, defining multiple control points sets and just one security guard. The results of times and distances calculated by the users and the system are shown on Figure 3.

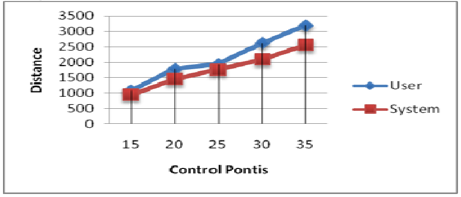


Figure 3.Distance calculated for one security guard and multiple control points sets

The system provides optimized calculations, so the time and distance are reduced. A complete working day shift can be fixed according the system results, for example, if the route calculated is too long or the time exceeds eight working hours, a new guard must be incorporated. The usage of a CBP-BDI agent allows the system to increase its performance since the ANN facilitates automatic route's calculation. The plans are more suitable to the user's skills because the planner takes into account their profiles and the results obtained in previous experiences, so they have a more realistic estimation of times to go between control points. Moreover, the CBP-BDI allows reducing the number of preplanning in the system. The planner with ANN only calculates the routes when it is necessary to replan or the system doesn't have any similar case in the memory. The administrator is able to redistribute the control point whenever he wonders. In this way the system avoids over-learning. In Figure 4b it is possible to see how the percentage of variation for the routes related to the increase the weeks. In Figure 4a shows the average number of estimated security guards needed to cover an entire area, which consisted on a mesh from 20 to 100 control points, with an increment of 5 control points. The results are clear, for example, for 80 control points, the users estimated 4 security guards, but the system recommended only 3.

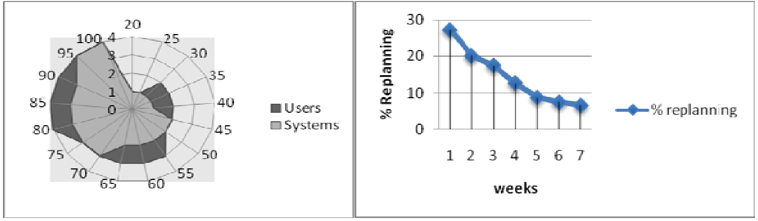


Figure 4(a). Average number of estimated security guards. (b) Percentage of replanning

The results obtained so far are positive. It is possible to determine the number of security guards needed to cover an entire area and the loops in the routes, so the human resources are optimized. In addition, the system provides the supervisors relevant information to monitor the workers activities, detecting incidences in the surveillance routes automatically and in real-time. The system presented can be easily adapted to other scenarios with similar characteristics.

Acknowledgements

This development has been supported by the projects MEC TIN2006-14630-C03-03 and JCYL-2002-05 (SA104A05). Special thanks to Telefónica Móviles and Sokymat for the technology provided.

6 References

[1] INOLOGY, PRESS NOTE, *Available: http://www.controldetiempos.com/sala_de_prensa.htm#absentismo.* (2005)

[2] C ANGULO AND R TELLEZ, *Distributed Intelligence for smart home appliances,* Tendencias de la minería de datos en España, Red Española de Minería de Datos, 2004.

[3] SOKYMAT.: SOKYMAT. *http://www.sokymat.com. (2006)*

[4] P RIGOLE, T HOLVOET AND Y BERBERS, *Using Jini to integrate home automation in a distributed software-system,* Computational Sciences Department, Belgium, (2002).

[5] GLEZ-BEDIA, M., CORCHADO, J.M.: *A planning strategy based on variational calculus for deliberative agents.* Computing and Information Systems Journal. Vol.10(1) 2-14. (2002)

[6] S GARFINKEL AND B ROSENBERG, *RFID: Applications, security, and privacy,* Addison-Wesley Professional, (2005), pp. 15-36.

[7] CORCHADO J. M. AND LAZA R.. *Constructing Deliberative Agents with Case-based Reasoning Technology,* International Journal of Intelligent Systems. Vol 18, No. 12, December. (2003) pp.: 1227-1241

[8] M WOOLDRIDGE AND N R JENNINGS, *Agent Theories, Architectures, and Languages: a Survey, In: Wooldridge and Jennings ed.,* Intelligent Agents, Springer-Verlag, (1995), pp. 1-22.

[9] KOLODNER J. *Case-Based Reasoning. Morgan Kaufmann (1993).*

[10] Q. MARTÍN, M T SANTOS AND Y DE PAZ, *Operations research: Resolute problems and exercises,* Pearson, (2005), pp. 189-190.

[11] N JENNINGS AND M WOOLDRIDGE, *Applications of Intelligent Agents, Queen Mary & Westfield College,* University of London, (1998).

[12] K S LEUNG, H D JIN AND Z B XU, *An expanding Self-organizing Neural Network for the Traveling Salesman Problem.* Neurocomputing, Vol. 62, (2004), pp. 267-292.

[13] BRATMAN, M.. *Intention, Plans and Practical Reason. Harvard U.P.,* Cambridge. (1987)

[14] ALLEN, J.F.: *Towards a General Theory of Action and Time. Artificial Intelligence* Vol. 23 (1984) 123-154

[15] SPALAZZI L. A *Survey on Case-Based Planning.* Artificial Intelligence Review, Vol. 16, Issue 1 (2001) 3-36

[16] POLLACK, M.E.: *The Uses of Plans. Artificial Intelligence.* Vol. 57 (1992) 43-68

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Numerical simulation of tornadoe like vortices around complex geometries

Frederique Drullion[1]

[1] *Department of Mathematics, Embry Riddle Aeronautical University*

emails: `drulliof@erau.com`

Abstract

This work is about numerical simulations of tornadoes. More precisely the problem is to, first, simulate, numerically, a tornado like vortex going through a populated area, and, then analyze the paths as it impacts facilities. This study started after a tornado classified F2 on the Fujita scale hit Embry- Riddle Aeronautical Daytona Beach campus in December 2006. The computational domain is the Florida ERAU campus and includes accurately all the different buildings. Computations of turbulent flow around several complex shaped buildings are performed using a $k - \epsilon$ turbulence model. They include a artificially tornado like vortex generated using a pressure gradient. The boundary conditions are discussed. The results obtained are physically reasonable.

Key words: CFD, Finite volume, $k - \epsilon$, Tornadoe like vortex.

Figures

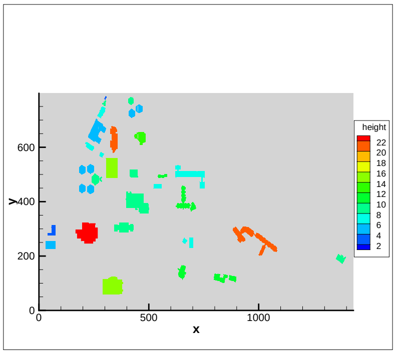


Figure 1: Computational domain: ERAU Daytona Beach campus

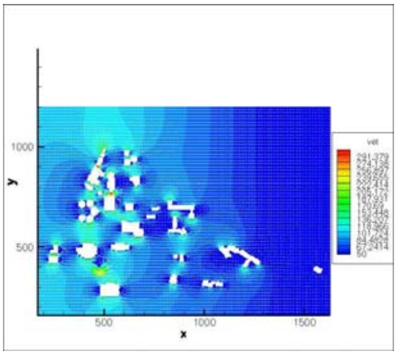


Figure 2: Flow simulation, horizontal West-East wind, tornadoe generated in the south west region.

Acknowledgements

This work has been supported by Embry Riddle Aeronautical University.

References

[1] X. Ding, *Physically based Simulation of Tornadoes*, University of Waterloo ON, CA, 2004.

[2] H. Bluestein, J. Golden, *A review of tornado observations. In The Tornado: its structure, dynamics, prediction, and hazards*, Geophysical Monograph **79** (1993) 319–352.

[3] W.Lewellen, *Tornado vortex theory.In The Tornado: its structure, dynamics, prediction, and hazards*, Geophysical Monograph **79** (1993) 19–39.

[4] B.Fiedler, *Numerical simulation of axisymmetric tornadogenesis in forced convection.*, Geophysical Monograph **79** (1993) 41–48.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

New pseudo-conform polynomial finite elements on quadrilaterals

Eric Dubach[1], Robert Luce[1] and Jean-Marie.Thomas[1]

[1] *Laboratoire de Mathématiques Appliquées, UMR 5142, INRIA, Université de Pau et des Pays de l'Adour, BP 1155, 64013 Pau Cedex, France.*

emails: eric.dubach@univ-pau.fr, Robert.Luce@univ-pau.fr,
Jean-marie.thomas@univ-pau.fr

Abstract

The aim of this paper is to develop new finite elements on quadrilaterals which give a polynomial approximation on each element of the triangulation. Pseudo-conform Lagrange and mixed finite elements are proposed.

Key words: Lagrange and mixed finite elements, polynomial approximation, non conform approximation, quadrilateral meshes

1 Introduction

Quadrilaterals and hexahedra are often used in meshers particularly in geophysical applications and in fluids mecanics. When the geometry and the medium are structured, regular rectangular meshes are used. Otherwise general convex quadrilaterals or hexahedra are used. Then, with isoparametric Lagrange finite elements([1],[2],[5]) or mixed finite elements ([3],[4]), we must construct finite elements on the mesh by using multilinear mappings to a reference rectangle or rectangular solid.
Lagrange finite elements do not converge on irregular quadrilaral or hexaedral meshes whereas Lagrange isoparametric finite elements do, but the jacobians of these mappings leads to non polynomial basis functions on the elements of the mesh . For mixte finite elements, consequences are even worse since the use of the Piola transform to work on the reference element is effective only when the mapping is linear otherwise a loss of order of convergence is observed ([6]).
To built our finite elements, we consider a quadrilateral as a distortion of a parallelogram and the Lagrange basis functions are built under conditions of weak-continuity of the unknowns between the elements. So the finite elements obtained are not conform but the conditions of weak-continuity are sufficient to ensure the expected order of convergence. In this study, we present finite elements of lower degree.

2 Geometry description

In this section we show that a distorded quadrilateral can be discribed with (by) a vector of distortion d as shown in 1. Let K be a convex quadrilateral of $\mathbb{R}^2$, let $a_i (1 \leq i \leq 4)$ denote its vertices and $\gamma_m (1 \leq m \leq 4)$ its edges. b_m is the edge midpoint of γ_m and $a_0 = \frac{1}{4}\sum_{1\leq i\leq 4} a_i$ is the center of K (isobarycenter of the vertices).
Let $\widehat{K} = [-1, +1]^2$ be the reference square. The vertices of $\widehat{K}$ are denoted by $\widehat{\mathbf{a}}_i,\ 1 \leq i \leq 4$ and its edges by $\widehat{\gamma}_m,\ 1 \leq m \leq 4$.
Let $F^\sharp$ be the invertible affin transform of $\mathbb{R}^2$ into $\mathbb{R}^2$ defined by $F^\sharp(0) = a_0,\ F^\sharp(\widehat{b}_m) = b_m \qquad m = 1, 2$, the image by $F^\sharp$ of the square $\widehat{K}$ is a parallelogram $K^\sharp$.

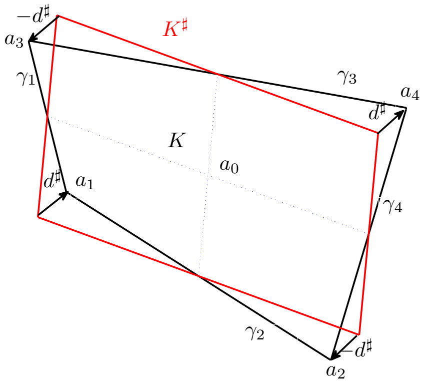


Figure 1: Vertices and edges numerotation and distortion parameter

Proposition 1

$$K\,\text{convex} \Longleftrightarrow \left|d_1^\sharp\right| + \left|d_2^\sharp\right| < 1$$

3 The model problem and the patch tests

We consider the following second order elliptic problem:

$$\begin{array}{rcl} -div(A\nabla u) & = & f \qquad \text{in}\,\Omega \\ u & = & 0 \quad \text{on}\,\Gamma \end{array} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded domain with a Lipschitz boundary $\Gamma = \partial\Omega$ and $A = (a_{i,j})$ a symmetric matrix of coefficients sufficiently smooth, satisfying

$$\forall x \in \bar{\Omega}, \forall \xi \in \mathbb{R}^d \quad c\sum_{i=1}^{d} \xi_i^2 \leq \sum_{i,j=1}^{d} a_{i,j}(x)\xi_i\xi_j \leq c^{-1}\sum_{i=1}^{d} \xi_i^2$$

Let τ_h be the triangulation of Ω into quadrilateral. We look for non conforming finite elements so the study of the consistancy ([7]) error gives the conditions that the basis functions must satisfy.

In this section we develop a variational formulation for the primal and mixed problem and give sufficient conditions on the spaces of discretisation to control the error of consistancy with the expected order (patch tests).

For Lagrange finite elements the patch test is satisfied if the mean value of the approximation on each edge is continuous.

For mixed finite elements the patch test is satisfied if the first momentum of the approximation on each edge is continuous.

4 Polynomial finite elements

This section is devoted to the construction of polynomial finite elements on quadrilaterals satisfying the patch test. In the first subsection we study the case of Lagrange finite elements and in the second the case of Raviart-Thomas finite elements. For each finite element we give explicitly the basic functions and local error estimates.

4.1 Lagrange finite elements

The purpose is to build a finite element (K, V_K, S_K) where the degrees of freedom are the vertices set of K, and V_K is a polynomial space.

Let be $V_K = \left\{q \in Q_2^K \cap P_3;\ q(b_m) = \frac{1}{2}\sum_{a_i \in \gamma_m} q(a_i),\ \text{for all}\ m = 1, ..., 4\right\}$ and $S_K = \{a_i; 1 \leq i \leq 4\}$

Theorem 2 *For any convex quadrilateral K, the triad (K, V_K, S_K) is a Lagrange finite element.*

4.2 Raviart-Thomas finite elements

Let us consider now the following vectorial polynomial space:

$$\Psi_K = \Big\{\mathbf{w} \in BDM_{[1]}^K;\ \text{for}\ 1 \leq m \leq 4\,, \forall p \in P_1 \\ \textstyle\int_{\gamma_m} p\mathbf{w}.\mathbf{n}\, d\sigma = \frac{1}{|\gamma_m|}\int_{\gamma_m} p\, d\sigma \int_{\gamma_m} \mathbf{w}.\mathbf{n}\, d\sigma\Big\}$$

and the set of degrees of freedom $\Sigma_K = \left\{\mathbf{w} \mapsto \int_{\gamma_m} \mathbf{w}.\mathbf{n}\, d\sigma; 1 \leq m \leq 4\right\}$. We have the following result:

Theorem 3 *For any convex quadrilateral K, the triad (K, Ψ_K, Σ_K) is a Raviart-Thomas finite element.*

5 Numerical tests

In this section some numerical tests illustrating the previous results are presented.

6 Some extensions

In this section, we give some explanations to built finite elements of higher order and we discuss the possibilities of extensions to 3D case.

References

[1] P.. G. Ciarlet, *The Finite Element Method for Elliptic Problems* Classics in Applied Mathematics, Vol.40, SIAM, Philadelphia, 2002. First edition published by North-Holland, Amsterdam, 1978.

[2] P.. G. Ciarlet, *Basic error estimates for elliptic problems*, in "Handbook of Numerical Analysis, Vol.II" (eds. P.G. Ciarlet and J.L. Lions), Elsevier, (1991), 17-351.

[3] J. E. Roberts, J.-M. Thomas *Mixed and hybrid methods.* In Handbook of Numerical Analysis, VOL.II, Finite Element Methods (Part 1), Elsevier Science Publishers B.V. (North-Holland), Amsterdam (1991) (1977) 523-639,.

[4] F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.

[5] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall,Englewood Cliffs,NJ, 1973

[6] D. N. Arnold, D. Boffi, R. S. Falk, *Approximation by quadrilateral finite elements*, Maths of Comp., **239** (2002), 909 - 922.

[7] S. C. Brenner L. R. Scott, *The Mathematical Theory of Finite Element Methods* Series: Texts in Applied Mathematics , Vol. 15 2nd ed., 2002

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Generic intersection of orthogonal groups

R. Durán Díaz[1], **L. Hernández Encinas**[2], **J. Muñoz Masqué**[2] **and A. Queiruga Dios**[3]

[1] *Dpt. Automatic, University of Alcalá de Henares*

[2] *Dpt. Information Processing and Coding, Applied Physics Institute, CSIC*

[3] *Dpt. Applied Mathematics, University of Salamanca*

emails: raul.duran@uah.es, luis@iec.csic.es, jaime@iec.csic.es, queirugadios@usal.es

Abstract

Let V be a finite-dimensional complex vector space and let $g, h\colon V \times V \to \mathbb{C}$ be two non-degenerate symmetric bilinear forms. Let G, H be the groups of isometries of g, h, respectively. If the endomorphism $L\colon V \to V$ associated to g, h is diagonalizable, then $\dim(G \cap H) = \sum_{i=1}^{r} \binom{m_i}{2}$, where m_i, $i = 1, \dots, r$, are the dimensions of the eigenspaces of L.

Key words: Bilinear form, Complex vector space, Diagonalizable endomorphism, Group of isometries.
MSC 2000: AMS codes (optional)

1 Introduction and preliminaries

Let V, W be two complex vector spaces of finite dimension and let $\mathcal{L}(V, W)$ be the space of $\mathbb{C}$-linear mappings from V into W. We write $\mathfrak{gl}(V) = \mathcal{L}(V, V)$ and we denote by $GL(V)$ the linear group of V, i.e., the group of invertible elements in $\mathfrak{gl}(V)$.

An element $A \in \mathfrak{gl}(V)$ is said to be an *isometry* of a symmetric bilinear form $g\colon V \times V \to \mathbb{C}$ if the following equation holds:

$$g\left(A(x), A(y)\right) = g(x, y), \quad \forall x, y \in V. \tag{1}$$

As a simple computation shows, we have

Lemma 1 *Let $g\colon V \times V \to \mathbb{C}$ be a symmetric bilinear form on an n-dimensional complex vector space V and let V', V'' be vector subspaces such that, (1) $g|_{V'}$ is non-degenerate, (2) $g(v, v'') = 0$, $\forall v \in V, \forall v'' \in V''$, and (3) $V = V' \oplus V''$.*

ISBN: 978-84-612-1982-7

Then, every isometry $A \in \mathfrak{gl}(V)$ of g can be written as

$$A = \begin{pmatrix} A' & \mathrm{O} \\ B & C \end{pmatrix}, \quad B \in \mathcal{L}(V', V''), \; C \in \mathfrak{gl}(V''),$$

and A' is an isometry of $g|_{V'}$.

Consequently, the structure of the set of isometries of a degenerate symmetric bilinear form g can be recovered from the non-degenerate part of g. Because of this, below we confine ourselves to consider only non-degenerate symmetric bilinear forms. In this case, every isometry of g is invertible, as the equation (1) implies $\det A = \pm 1$, and the set of all isometries of g is a subgroup of $GL(V)$, which is denoted by G. By choosing a orthonormal basis in V, every element of G is represented by an orthogonal matrix and we have an isomorphism $G \cong O(n, \mathbb{C})$.

We also remark on the fact that G is a closed subgroup in $GL(V)$ and hence, G is a Lie subgroup of the linear group of V, which Lie algebra will be denoted by $\mathfrak{g}$.

2 Main result

Theorem 2 *Let V be an n-dimensional complex vector space and let*

$$g, h \colon V \times V \to \mathbb{C}$$

be two symmetric bilinear forms, which are assumed to be non-degenerate. Let G, H be the groups of isometries of g, h, respectively and let $L \colon V \to V$ be the endomorphism associated to g, h, i.e., $g(x, L(y)) = h(x, y)$, $\forall x, y \in V$. If L is diagonalizable, then

$$\dim(G \cap H) = \sum_{i=1}^{r} \binom{m_i}{2},$$

where m_i, $i = 1, \ldots, r$, are the dimensions of the eigenspaces of L.

Sketch of the proof. Let α_i, $i = 1, \ldots, r$, be the distinct eigenvalues of L and let $E(\alpha_i)$ be the eigenspace attached to α_i. As L is diagonalizable, we have $V = \oplus_{i=1}^{r} E(\alpha_i)$ and $E(\alpha_i)$ and $E(\alpha_j)$ are orthogonal with respect to both metrics for $i \neq j$. There exist basis of every subspace $E(\alpha_i)$ to which the Gram-Schmidt process can be applied. Collecting all theses bases, we obtain a basis $(v_1, \ldots, v_n)$ of eigenvectors for L which is also g-orthonormal and the matrices of g and h in this basis are,

$$\begin{aligned} &M_g = I_n = n \times n \text{ identity matrix}, \\ &M_h = \text{diagonal}\left(\alpha_1, \overset{(m_1}{\ldots}, \alpha_1, \ldots, \alpha_r, \overset{(m_r}{\ldots}, \alpha_r\right), \; m_1 + \ldots + m_r = n. \end{aligned}$$

Let $\mathfrak{g}$ (resp. $\mathfrak{h}$) be the Lie algebra of G (resp. H). As is known ([2, Theorem 3.31]) the exponential map $\exp \colon \mathfrak{g} \to G$ induces an diffeomorphism from an open neighbourhood of the origin in $\mathfrak{g}$ onto an open neighbourhood of the unit element in G. Hence

$\dim(G \cap H) = \dim(\mathfrak{g} \cap \mathfrak{h})$, and we are led to determine the Lie algebra of the intersection subgroup. As $\mathfrak{g} = \{A \in \mathfrak{gl}(V) : g\,(x, A(y)) + g\,(A(x), y) = 0, \forall x, y \in V\}$, and similarly for $\mathfrak{h}$, we conclude that $\mathfrak{g} \cap \mathfrak{h}$ can be identified to the subspace of $n \times n$ skew-symmetric matrices $A = (a_{ij})$ such that, $A^t M_h + M_h A = 0$. By decomposing A in blocks,

$$A = \begin{pmatrix} A_{11} & \dots & A_{1r} \\ \vdots & \ddots & \vdots \\ A_{r1} & \dots & A_{rr} \end{pmatrix},$$

each A_{ij} being a $m_i \times m_j$ matrix for $i, j = 1, \dots, r$, we obtain $A_{ij} = 0, i \neq j$, and the submatrices $A_{11}, \dots, A_{rr}$ are arbitrary. As $\dim \mathfrak{o}(m, \mathbb{C}) = \binom{m}{2}$, we can conclude. □

Taking [1, Chapter 7,Theorem 1] into account, we also obtain

Corollary 3 *Let $\mathcal{U} \subset S^2 V^*$ be the subset of non-degenerate bilinear forms. The pairs $(g, h) \in \mathcal{U} \times \mathcal{U}$ for which the conclusion of the theorem above holds is a dense subset in $\mathcal{U} \times \mathcal{U}$.*

3 Concluding remarks

Remark 4 *According to the proof of the previous theorem, the matrices of the form*

$$\exp(\tilde{A}_{11}) \cdots \exp(\tilde{A}_{rr}), \quad A_{ii} \in \mathfrak{o}(m_i, \mathbb{C}),\ 1 \leq i \leq r,$$

$$\tilde{A}_{ii} = \begin{pmatrix} \mathrm{O}_{\mu_i, \mu_i} & \mathrm{O}_{\mu_i, m_i} & \mathrm{O}_{\mu_i, n-\mu_{i+1}} \\ \mathrm{O}_{m_i, \mu_i} & A_{ii} & \mathrm{O}_{m_i, n-\mu_{i+1}} \\ \mathrm{O}_{n-\mu_{i+1}, \mu_i} & \mathrm{O}_{n-\mu_{i+1}, m_i} & \mathrm{O}_{n-\mu_{i+1}, n-\mu_{i+1}} \end{pmatrix},$$

where $\mu_i = m_1 + \ldots + m_{i-1}$, and $\mathrm{O}_{\mu,\nu}$ denotes the null $\mu \times \nu$ matrix, span the intersection group $G \cap H$. Hence the problem of computing the intersection group is feasible: In fact, it reduces to exponentiate skew-symmetric matrices of size $m_1, \dots, m_r$.

Remark 5 *The previous theorem is no longer true if the endomorphism L is not diagonalizable. For example, for the metrics g, h with matrices*

$$M_g = \begin{pmatrix} \overbrace{\begin{pmatrix} 0 & \dots & 1 \\ \vdots & \iddots & \vdots \\ 1 & \dots & 0 \end{pmatrix}}^{(k} & \mathrm{O} \\ \mathrm{O} & \overbrace{\begin{pmatrix} 0 & \dots & 1 \\ \vdots & \iddots & \vdots \\ 1 & \dots & 0 \end{pmatrix}}^{(n-k} \end{pmatrix},$$

$$M_h = \left(\begin{array}{cc} \overbrace{\begin{pmatrix} 0 & 0 & \dots & 1 & \alpha \\ 0 & 0 & \dots & \alpha & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 1 & \alpha & \dots & 0 & 0 \\ \alpha & 0 & \dots & 0 & 0 \end{pmatrix}}^{(k} & \mathrm{O} \\ \mathrm{O} & \overbrace{\begin{pmatrix} 0 & 0 & \dots & 1 & \alpha \\ 0 & 0 & \dots & \alpha & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 1 & \alpha & \dots & 0 & 0 \\ \alpha & 0 & \dots & 0 & 0 \end{pmatrix}}^{(n-k} \end{array}\right),$$

respectively, as a computation shows, we have $\dim(\mathfrak{g} \cap \mathfrak{h}) = \min(k, n-k)$*, whereas* α *is the only eigenvalue of* L *and* $\dim E(\alpha) = 2$*.*

4 An example

Assume $\dim V = n = 5$, and that L has two distinct eigenvalues α, β such that $\dim E(\alpha) = 2$, $\dim E(\beta) = 3$. In this case, $\mathfrak{g} \cap \mathfrak{h}$ is identified to the matrices of the form

$$A = \begin{pmatrix} A_{11} & O \\ O & A_{22} \end{pmatrix}, \qquad A_{11} = \begin{pmatrix} 0 & d \\ -d & 0 \end{pmatrix}, A_{22} = \begin{pmatrix} 0 & a & b \\ -a & 0 & c \\ -b & -c & 0 \end{pmatrix}.$$

According to Remark 4, the intersection group is generated by $\exp \tilde{A}_{11} \exp \tilde{A}_{22}$. Exponentiating, we obtain

$$\exp \tilde{A}_{11} \exp \tilde{A}_{22} = \begin{pmatrix} \begin{pmatrix} \cos d & \sin d \\ -\sin d & \cos d \end{pmatrix} & O \\ O & \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \end{pmatrix} \end{pmatrix},$$

where $v = (a, b, c)$, and

$$\begin{aligned}
\lambda_{11} &= \frac{c^2 + \left(a^2 + b^2\right)\cos(|v|)}{|v|^2}, \\
\lambda_{12} &= \frac{a|v|\sin(|v|) + bc\,(\cos(|v|) - 1)}{|v|^2}, \\
\lambda_{13} &= \frac{b|v|\sin(|v|) - ac\,(\cos(|v|) - 1)}{|v|^2}, \\
\lambda_{21} &= -\frac{a|v|\sin(|v|) - bc\,(\cos(|v|) - 1)}{|v|^2}, \\
\lambda_{22} &= \frac{b^2 + \left(a^2 + c^2\right)\cos(|v|)}{|v|^2}, \\
\lambda_{23} &= \frac{c|v|\sin(|v|) + ab\,(\cos(|v|) - 1)}{|v|^2}, \\
\lambda_{31} &= -\frac{b|v|\sin(|v|) + ac\,(\cos(|v|) - 1)}{|v|^2}, \\
\lambda_{32} &= -\frac{c|v|\sin(|v|) - ab\,(\cos(|v|) - 1)}{|v|^2}, \\
\lambda_{33} &= \frac{a^2 + \left(b^2 + c^2\right)\cos(|v|)}{|v|^2}.
\end{aligned}$$

References

[1] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[2] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Company, London, 1971.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2008
La Manga, Spain, 13-16 June 2008

Models with time-dependent parameters using transform methods: application to Heston's model

A. Elices[1]

[1]*Senior Quant member of the model validation group at Grupo Santander.*
Email: aelices@gruposantander.com

Abstract

This paper presents a methodology to introduce time-dependent parameters for a wide family of models preserving their analytic tractability. This family includes hybrid models with stochastic volatility, stochastic interest-rates, jumps and their non-hybrid counterparts. The methodology is applied to Heston's model. A bootstrapping algorithm is presented for calibration. A case study works out the calibration of the time-dependent parameters to the volatility surface of the Eurostoxx 50 index. The methodology is also applied to the analytic valuation of forward start vanilla options driven by Heston´s model. This result is used to explore the forward skew of the case study.

Key workds— Smile, forward skew, hybrid models, transform methods, Heston's model, piecewise constant parameters, characteristic function.

1 Introduction

Considering the smile effect when pricing and hedging exotics is an issue of major concern for both traders and practitioners. Pricing complex exotic derivatives according to a particular skew model is typically carried out using Monte-Carlo methods. However, calibration of model parameters to market involves many evaluations of vanilla products. Therefore, an analytic or at least a quick evaluation method for vanilla options is crucial. One of the major drawbacks of analytic models is that they depend on just a few parameters which do not provide enough degrees of freedom to fit the market at several maturities. The whole motivation of this paper is to provide more degrees of freedom by introducing piecewise constant time-dependent parameters.

This problem is not new and several authors have already given solutions for specific models. The main contribution of this paper is a methodology to extend not only a specific model, but a wide family of them so that time-dependent parameters can be introduced preserving analytic tractability. This family includes models with stochastic volatility (e.g. [Heston, 1993]), jumps (e.g. [Merton, 1976]) and hybrid models with stochastic volatility and stochastic interest rates correlated with the underlying (e.g. [Bakshi, 1997] and [Scott, 1997]). The methodology is based on characteristic function methods which describe the probability distributions of the stochastic processes in terms of the characteristic function. This general methodology is illustrated by applying it to Heston's model for valuation of both spot and forward start vanilla options. This same problem has already been addressed by [Mikhailov et al, 2005] from a completely different perspective using partial differential equations. In the latter paper, the solution to Heston´s partial differential equation is obtained for successive periods with different sets of parameters. The solution of the last period is used as initial condition for the preceding one. The solution of this period is applied to the preceding one until the first period is reached.

Another very interesting related work may be found in [Piterbarg, 2005] and [Piterbarg, 2006]. These papers derive approximate formulas to imply time-dependent parameters in between maturities from average parameters which fit market prices at each maturity. This allows an independent calibration of average parameters to fit market prices at each maturity (this data may already be available in trading desks). From these average parameters, a time-dependent model is implied. This model replicates the distribution and market prices at each maturity. The specific model used to describe the underlying process is the displaced diffusion stochastic volatility model in [Andersen et al, 2002]. A local volatility function controls the slope of the implied volatility smile allowing independent Brownian motions for the stochastic volatility and the underlying process.

[Britten-Jones et al, 2000] prove that for all continuous processes the expected value of realized variance up to a given maturity is defined by vanilla option prices for that maturity with respect to a continuum of strikes. This means that all possible models which calibrate vanilla option prices must have the same expected value of realized variance. For instance, Dupire's local volatility model fits the expected value of realized variance by a deterministic function (the local volatility). It is important to have in mind that exotic pricing has to be taken with care and prices given by any model should be considered in its right context. [Schoutens et al, 2004] or [Britten-Jones et al, 2000] give some examples of how a variety of different models fitting the smile option prices give considerably different prices depending on the hypothesis of the underlying process. Section 5 works out two different calibration sets which fit market prices. They will be compared in section 7 in terms of the forward skew and the price differences will be explained. Most traders and practitioners like to compare the prices of different models and trade and hedge with the model which gives the closest price to market (or what they think that should be the correct market price).

The paper is organized in five sections. Section 2 (the main contribution of the paper) presents a general methodology to derive characteristic functions for a time horizon where the parameters of the underlying process may change. This characteristic function is expressed in terms of the characteristic functions of each sub-period where parameters change. Section 3 applies the methodology to the well-known model by Heston and section 4 proposes a bootstrapping calibration algorithm. Section 5 presents a case study which works out two different calibration sets of a Heston model with time-dependent parameters to the volatility surface of the Eurostoxx 50 index. Section 6 applies again the methodology to derive a semi-analytic formula for the valuation of forward start vanilla options driven by Heston´s model. Section 7 uses this formula to explore the forward skew of both calibration sets provided by section 5 and explaing why they give different results. The paper ends with some conclusions. Appendix A derives a more general version of Heston's characteristic function so that the new methodology can be applied.

2 Characteristic functions of models with time dependent parameters

Consider the characteristic function (1) of the distribution of a Markov *N*-dimensional process $\mathbf{x}(t)=(x_1(t),\cdots,x_N(t))$. From a mathematical point of view, the characteristic function is the Fourier Transform of the density function.

$$\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)=\mathbf{E}\left(e^{i\mathbf{X}\cdot\mathbf{x}_v}\right)=\int_{\mathbf{R}^N} e^{i\mathbf{X}\cdot\mathbf{x}_v} f_{uv}(\mathbf{X}/\mathbf{x}_u)d\mathbf{X} \tag{1}$$

The notation $\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$ and $f_{uv}(\mathbf{X}/\mathbf{x}_u)$ refers to the characteristic function and the density function of the joint distribution of the process $\mathbf{x}_v=\mathbf{x}(t_v)$ at time t_v, conditioned by its initial value $\mathbf{x}_u=\mathbf{x}(t_u)$ at t_u. $\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$ is a function of the vector $\mathbf{X}=(X_1,\cdots,X_N)$ and the notation $\mathbf{X}\cdot\mathbf{x}_v=\sum X_k x_k(t_v)$ refers to the inner product of vectors **X** and $\mathbf{x}_v$.

Consider now the family of exponential characteristic functions of the form (2) with exponents linear in the stochastic processes $\mathbf{x}_u$ at time t_u. The vector function $\mathbf{D}_{uv}(\mathbf{X})=(D_{uv,1}(\mathbf{X}),\cdots,D_{uv,N}(\mathbf{X}))$ and the function $C_{uv}(\mathbf{X})$ depend not only on **X**, but on the parameters of the particular model under consideration in the period from t_u to t_v. This parameter dependence is dropped to simplify notation.

$$\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)=\exp\left(C_{uv}(\mathbf{X})+\mathbf{D}_{uv}(\mathbf{X})\cdot\mathbf{x}_u\right) \tag{2}$$

A wide range of models belong this class. Examples are provided below. Equation (3) presents the form of the characteristic function of Merton's lognormal jump diffusion model [Merton, 1976]. The variable g_u is the sum of all jumps (which happen randomly following a Poisson process) up to time t_u.

$$\varphi_{uv}(G/g_u)=\exp\left(C_{uv}(G)+iGg_u\right) \tag{3}$$

The form of the characteristic function of the fixed income short rate model by Cox, Ingersoll and Ross [Cox et al, 1985a] is given by equation (4), where r_u is the short rate at time t_u.

$$\varphi_{uv}(R/r_u) = \exp\left(C_{uv}(R) + iD_{uv}(R)r_u\right) \tag{4}$$

Similarly, the form of the characteristic function of Heston's lognormal model with stochastic volatility [Heston, 1993] is presented by equation (5), where x_u is the logarithm of the underlying stock and v_u is the variance (both at time t_u).

$$\varphi_{uv}(X,V/x_u,v_u) = \exp\left(C_{uv}(X,V) + D_{uv}(X,V)v_u + iXx_u\right) \tag{5}$$

An example of a hybrid model combining the latter four altogether such as [Bakshi, 1997] or [Scott, 1997] is given by equation (6), where *C* and *D* depend on the four variables *X*, *V*, *R* and *G*.

$$\varphi_{uv}(X,V,R,G/x_u,v_u,r_u,g_u) = e^{C_{uv}+D_{uv,2}r_u+D_{uv,1}v_u+Xx_u+Gg_u} \tag{6}$$

All these models allow for quick semi-analytic formulas to price vanilla options by using transform methods that invert the characteristic function. The main drawback of these models is that they depend on a few parameters which do not provide enough degrees of freedom to calibrate the market. The goal of this section is to derive the characteristic function of a process that may have time- dependent parameters.

As vanilla options used for calibration are usually European, all the information of a Markov process with independent increments up to an instant is given by the joint probability distribution of the stochastic variables which describe it at that instant (marginal distributions can always be calculated from the joint distribution). In addition, all the information necessary to continue the evolution of this process from an instant t_u to a later one t_v, is the joint distribution at t_u and the evolution law from t_u to t_v. Fig. 1 represents graphically this information: from 0 to t_u the process is described by the characteristic function $\varphi_{0u}(\mathbf{X}/\mathbf{x}_0)$ conditioned by $\mathbf{x}_0$ and from t_u to t_v, the process is described by $\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$ conditioned by $\mathbf{x}_u$. Note that both characteristic functions may represent the underlying evolution with different parameters. In this context, the goal is to obtain the characteristic function $\varphi_{0v}(\mathbf{X}/\mathbf{x}_0)$ of the joint distribution at t_v given $\mathbf{x}_0$ in terms of $\varphi_{0u}(\mathbf{X}/\mathbf{x}_0)$ and $\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$.

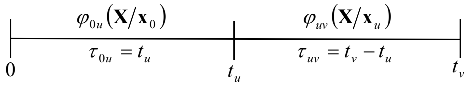

$\varphi_{0u}(\mathbf{X}/\mathbf{x}_0)$
$\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$
$\tau_{0u} = t_u$
$\tau_{uv} = t_v - t_u$
0
t_u
t_v

Fig. 1: Graphical representation of a Markov process with independent increments in two consecutive periods.

Equation (7) shows the definition of the characteristic function under search.

$$\varphi_{0v}(\mathbf{X}/\mathbf{x}_0) = \int_{\mathbf{R}^N} d\mathbf{x}_v e^{i\mathbf{X}\cdot\mathbf{x}_v} f_{0v}(\mathbf{x}_v/\mathbf{x}_0) \tag{7}$$

As non-overlapping intervals are independent, the density from 0 to t_v is the product of the densities from 0 to t_u and from t_u to t_v, summed over all intermediate paths $\mathbf{x}_u$ as shown in equation (8).

$$f_{0v}(\mathbf{x}_v/\mathbf{x}_0) = \int_{\mathbf{R}^N} d\mathbf{x}_u f_{0u}(\mathbf{x}_u/\mathbf{x}_0) f_{uv}(\mathbf{x}_v/\mathbf{x}_u) \tag{8}$$

Substituting (8) in (7) and exchanging the sum order (summing over x_v first) yields equation (9):

$$\varphi_{0v}(\mathbf{X}/\mathbf{x}_0) = \int_{\mathbf{R}^N} d\mathbf{x}_u f_{0u}(\mathbf{x}_u/\mathbf{x}_0) \int_{\mathbf{R}^N} d\mathbf{x}_v e^{i\mathbf{X}\cdot\mathbf{x}_v} f_{uv}(\mathbf{x}_v/\mathbf{x}_u) \tag{9}$$

As the second integral of (9) is the definition of $\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$, equation (9) becomes (10):

$$\varphi_{0v}(\mathbf{X}/\mathbf{x}_0) = \int_{\mathbf{R}^N} d\mathbf{x}_u f_{0u}(\mathbf{x}_u/\mathbf{x}_0) \varphi_{uv}(\mathbf{X}/\mathbf{x}_u) \tag{10}$$

Substituting equation (2) in equation (10), applying the definition of $\varphi_{0u}(\mathbf{X}/\mathbf{x}_0)$ in (12) and substituting $\varphi_{0u}(\mathbf{X}/\mathbf{x}_0)$ according to equation (2) in (13) yields:

$$\varphi_{0v}(\mathbf{X}/\mathbf{x}_0) = \int_{\mathbf{R}^N} d\mathbf{x}_u f_{0u}(\mathbf{x}_u/\mathbf{x}_0) \exp\left(C_{uv}(\mathbf{X}) + \mathbf{D}_{uv}(\mathbf{X})\cdot\mathbf{x}_u\right) \tag{11}$$

$$= \exp\left(C_{uv}(\mathbf{X})\right) \int_{\mathbf{R}^N} d\mathbf{x}_u f_{0u}(\mathbf{x}_u/\mathbf{x}_0) \exp\left(i\left(i^{-1}\mathbf{D}_{uv}(\mathbf{X})\right)\cdot\mathbf{x}_u\right) \tag{12}$$

$$= \exp\left(C_{uv}(\mathbf{X})\right)\varphi_{0u}\left(i^{-1}\mathbf{D}_{uv}(\mathbf{X})/\mathbf{x}_0\right) \tag{13}$$

$$= \exp\left(C_{uv}(\mathbf{X}) + C_{0u}\left(i^{-1}\mathbf{D}_{uv}(\mathbf{X})\right) + \mathbf{D}_{0u}\left(i^{-1}\mathbf{D}_{uv}(\mathbf{X})\right)\cdot\mathbf{x}_0\right) \tag{14}$$

Applying equation (2) to the interval 0 to t_v yields (15):

$$\varphi_{0v}(\mathbf{X}/\mathbf{x}_0) = \exp\left(C_{0v}(\mathbf{X}) + \mathbf{D}_{0v}(\mathbf{X})\cdot\mathbf{x}_0\right) \tag{15}$$

Identifying terms between equations (14) and (15) yields (16): the expression of $\varphi_{0v}(\mathbf{X}/\mathbf{x}_0)$ in terms of $\varphi_{0u}(\mathbf{X}/\mathbf{x}_0)$ and $\varphi_{uv}(\mathbf{X}/\mathbf{x}_u)$.

$$\begin{cases} C_{0v}(\mathbf{X}) = C_{uv}(\mathbf{X}) + C_{0u}\left(i^{-1}\mathbf{D}_{uv}(\mathbf{X})\right) \\ \mathbf{D}_{0v}(\mathbf{X}) = \mathbf{D}_{0u}\left(i^{-1}\mathbf{D}_{uv}(\mathbf{X})\right) \end{cases} \tag{16}$$

Consider now Fig. 2 with a series of periods in which the parameters of the process are different.

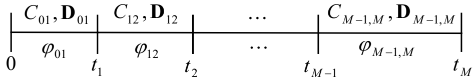


Fig. 2: Example of a five period process.

The characteristic function φ_{0M} at a given maturity t_M can be obtained recursively applying equation (16) to $\varphi_{M-1,M}$, given by equation (2), and $\varphi_{0,M-1}$. $\varphi_{0,M-1}$ is obtained applying again equation (16) to $\varphi_{M-2,M-1}$, given by equation (2), and $\varphi_{0,M-2}$. This process continues until φ_{02} is obtained applying equation (16) to φ_{01} and φ_{12} where both are calculated using equation (2).

The marginal distribution of the *h*th stochastic variable can be calculated from the definition of characteristic function by setting to zero all the X_k except for X_h as indicated in equation (17).

$$\varphi_{0v}(X_h/x_h(t_0)) = \varphi_{0v}\left((X_1,\cdots,X_N)/\mathbf{x}_0\right)\Big|_{X_k=0 \quad k\neq h} \tag{17}$$

3 Application to Heston's model

Equation (18) presents Heston´s underlying process S_t,

$$\begin{cases} dS_t = \mu S_t dt + \sqrt{v_t}S_t dW_t \\ dv_t = \kappa(\theta - v_t)dt + \sigma\sqrt{v_t}dY_t \end{cases} \qquad d\langle W_t, Y_t\rangle = \rho dt \tag{18}$$

where v_t is the stochastic variance. This variance follows an Ornstein-Uhlenbeck process where κ is the mean reversion rate, θ is the long term variance and σ is the volatility of the variance process. There is a correlation ρ between the Brownian motions which drive the underlying process and the variance. The parameter μ is the risk-neutral drift[2].

$$C = P(0,T)\mathbf{E}\left((S_T - K)^+\right) \tag{19}$$

Equation (19) shows the price of a call option where *K* is the strike price, $P(0,T)$ is the discount factor from expiry *T* to present time, the expectation is calculated with the information of present time and the measure of the expectation is the same used to calculate the risk neutral drift μ of equation (18).

$$C = P(0,T)\mathbf{E}\left(e^{x_T}\mathbf{1}_{\{x_T>\ln K\}}\right) - P(0,T)K\mathbf{E}\left(\mathbf{1}_{\{x_T>\ln K\}}\right) \tag{20}$$

Equation (20) breaks the price into two expectations and expresses the payoff in terms of $x_t = \ln S_t$. In order to calculate these expectations, it is necessary to know the marginal distribution of x_T given x_0 and v_0. Although the density function of this distribution is not analytic, Heston derived an analytic expression for the characteristic function [Heston, 1993]. In order to apply the methodology developed in section 2, the characteristic function of the joint distribution of x_T and v_T given x_0 and v_0 (and not the marginal) must be

[2] By risk-neutral drift it is meant the drift that forces the process S_t/N_t to be a martingale under the chosen numeraire N_t. Throughout this paper the numeraire N_t will be the bank account B_t=exp(r×t) Under these assumptions, μ=*r-q* where *r* is the risk free rate and *q* is the continuous dividend yield of the stock.

calculated. Appendix A presents the details of this calculation. The final result for the period from t_u to t_v is given by equation (21),

$$\varphi_{uv}(\mathbf{X}/\mathbf{x}_u) = e^{C_{uv}(\mathbf{X})+D_{uv,2}(\mathbf{X})v(t_u)+D_{uv,1}(\mathbf{X})x(t_u)} \tag{21}$$

where $\mathbf{X} = (X, V)$, $\mathbf{x}_u = (x(t_u), v(t_u))$. $C_{uv}(\mathbf{X})$ is given by equation (70) (with $C^0 = 0$ and $D^0 = iV$). $D_{uv,2}(\mathbf{X})$ is given by equation (63), $D_{uv,1}(\mathbf{X}) = iX$ and $\tau = t_v - t_u$. The variable X of the characteristic function corresponds to the logarithm of the underlying stock and V corresponds to the variance process.

$$\begin{cases} C_{0v}(X,V) = C_{uv}(X,V) + C_{0u}\left(X, i^{-1}D_{uv,2}(X,V)\right) \\ D_{0v,2}(X,V) = D_{0u,2}\left(X, i^{-1}D_{uv,2}(X,V)\right) \\ D_{0v,1}(X,V) = iX \end{cases} \tag{22}$$

If time dependent parameters across several periods are considered, equation (16) becomes (22) and the same procedure of section 2 can be applied to get the joint characteristic function (23) from 0 to any time.

$$\varphi_{0v}(X, V/x_0, v_0) = e^{C_{0v}(X,V)+D_{0v,2}(X,V)v_0+D_{0v,1}(X,V)x_0} \tag{23}$$

The marginal characteristic function of x_T is obtained by simply evaluating the joint characteristic function at $V = 0$ according to equation (24).

$$\varphi_{0T}(X/x_0, v_0) = \varphi_{0T}(X, 0/x_0, v_0) = \mathbf{E}\left(e^{iXx_T}\right) \tag{24}$$

Equation (25) presents the inversion formula to calculate the probability *P* from a distribution defined by its characteristic function φ. This formula is found in [Kendal, 1987] and [Shephard, 1991a]. A discussion on Fourier inversion formulas is also available in [Feller, 1971].

$$P(x > a) = \frac{1}{2} + \frac{1}{2\pi}\int_0^\infty \frac{1}{iX}\left(\frac{\varphi(X)}{e^{iXa}} - \frac{\varphi(-X)}{e^{-iXa}}\right)dX \tag{25}$$

When $\varphi(-X)$ and $\varphi(X)$ are complex conjugates equation (25) can be reduced to (26), where $\mathbf{Re}(.)$ stands for the real part (this is the final result given in [Heston, 1993]). It can be verified that this condition is satisfied for both equation (21) (flat model) and (22) (model with time dependent parameters).

$$P(x > a) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathbf{Re}\left(\frac{\varphi(X)e^{-iXa}}{iX}\right)dX \tag{26}$$

The second expectation in (20) is calculated using the inversion formula (25) with $\varphi(X) = \varphi_{0\mathrm{T}}(X/\mathbf{x}_0)$ and $a = \ln K$. The first expectation in (20) can be obtained similarly using the characteristic function $\tilde{\varphi}_{0\mathrm{T}}(X/\mathbf{x}_0)$ of equation (27).

$$\tilde{\varphi}_{0T}(X/\mathbf{x}_0) = \frac{\varphi_{0T}(X - i/\mathbf{x}_0)}{\varphi_{0T}(-i/\mathbf{x}_0)} = \frac{\mathbf{E}\left(e^{i(X-i)x_T}\right)}{\mathbf{E}\left(e^{x_T}\right)} = \frac{\mathbf{E}\left(e^{iXx_T + x_T}\right)}{\mathbf{E}\left(e^{x_T}\right)} \tag{27}$$

To apply the inversion formula (25), it is necessary that $\tilde{\varphi}_{0T}(0/\mathbf{x}_0)$ equals one, so that $\tilde{\varphi}_{0T}(0/\mathbf{x}_0)$ be a characteristic function of a fictitious density $\tilde{f}_{0T}(0/\mathbf{x}_0)$ that sums one over the real domain. Therefore, $\varphi_{0T}(X - i/\mathbf{x}_0)$ is normalized by the constant $\varphi_{0T}(-i/\mathbf{x}_0)$ (the forward price of the underlying).

$$\mathbf{E}\left(e^{x_T}\mathbf{1}_{\{x_T > \ln K\}}\right) = \mathbf{E}\left(S_T\right)\tilde{P}(x_T > \ln K) \tag{28}$$

The probability $\tilde{P}$ given by formula (25) with $\varphi(X) = \tilde{\varphi}_{0\mathrm{T}}(X/\mathbf{x}_0)$ and $a = \ln K$ will yield the desired expectation normalized by the forward price. Therefore the expectation will be given by equation (28) and the final call price by (29):

$$C = P(0,T)\mathbf{E}\left(S_T\right)\tilde{P}(x_T > \ln K) - P(0,T)KP(x_T > \ln K) \tag{29}$$

Fast and accurate methods to implement the inversion formula (25) can be found in [Shephard, 1991b] and [Davies, 1973]. A well-known algorithm to avoid the singularity at zero of (25) and apply the Fast Fourier Transform algorithm controlling the precision is described in [Lee, 2005] and [Carr et al, 1998].

4 Calibration

The marginal density function of the underlying at a given maturity is completely determined by a continuum of vanilla prices dependent on the strike. A good approximation of this distribution can be obtained from the interpolation of the implied volatilities of vanilla calls and puts with respect to the strike. As more exotic path-dependent options are not as liquid as vanillas and they are usually traded over the counter, the market provides quite limited information about the evolution of the underlying process in between maturities. In addition, introducing path-dependent products in the calibration is quite challenging because analytic solutions are usually not available. Therefore, any model calibrated to market should at least reproduce the marginal distribution of the underlying at the maturities for which vanilla products are quoted.

The most immediate and probably the simplest and quickest solution to calibrate a model with time-dependent parameters is a bootstrapping algorithm. The periods where the parameters change are given by the periods in between the maturities of the vanilla products used for calibration. Each period is calibrated independently starting from the first to the last solving a minimisation problem with the objective function of equation (30). The weights w_i are chosen to give the highest priority to the at the money (ATM) options. These weights will decrease as the moneyness of the option gets apart from ATM.

$$FO = \sum_{i=1}^{M} \frac{w_i}{\sum w_j}\left(price_i^{model} - price_i^{market}\right)^2 \tag{30}$$

When the parameters of the first period are calibrated to fit vanilla prices expiring on the first maturity, they are fixed. The parameters of the second period are then calibrated to fit vanilla prices expiring on the second maturity leaving the parameters of the first period fixed. Now, the parameters of the first two periods are fixed and the parameters of the third period are calibrated to fit vanilla prices expiring on the third maturity. This process continues until the last period. The advantage of a bootstrapping algorithm is that each period involves an optimisation with only the parameters of that period.

5 Case Study: Calibration of the Eurostoxx 50 index

The whole methodology proposed in this paper is applied to Heston´s model for the calibration of the Eurostoxx 50 index. The spot price is 3868.64€ and the volatility surface is given by Table 1. The leftmost column shows the moneyness of the options with respect to the spot price (the strikes are the moneyness times the spot).

TABLE 1: EUROSTOXX 50 VOLATILITY SURFACE.

K \ Mat	**1m**	**3m**	**6m**	**9m**	**1y**	**2y**	**3y**	**4y**	**5y**	**10y**
0.85	23.0	18.7	18.5	18.6	19.1	19.7	20.6	21.5	22.2	25.8
0.90	18.9	16.7	17.0	17.2	17.8	18.8	19.8	20.8	21.5	25.3
0.95	15.2	14.7	15.5	16.0	16.6	17.8	19.0	20.0	20.8	24.7
1.00	12.2	13.2	14.1	14.8	15.5	16.9	18.2	19.3	20.2	24.2
1.05	11.6	12.3	13.1	13.9	14.4	16.1	17.5	18.7	19.5	23.7
1.10	13.3	12.3	12.6	13.2	13.7	15.4	16.9	18.1	19.0	23.2
1.15	15.6	12.9	12.4	12.7	13.2	14.8	16.3	17.5	18.5	22.7

To avoid problems with discrete dividend payments, it is more convenient to calibrate the model of the forward price of the underlying F_t^P delivered on the last maturity date *P*, rather than the underlying spot S_t. When pricing exotics by Monte Carlo, the evolution of the forward is simulated and the spot price is recovered from the forward at each time using equation (31), where *NPV* is the net present value of all discrete dividends from *t* to the delivery date *P* of the forward.

$$F_t^P = \frac{S_t - NPV(dividends_{t-P})}{P(t,P)} \tag{31}$$

The prices of the vanilla options on S_t should also be replaced by equivalent vanilla options on F_t^P according to equation (32). If both interest rates and dividends are deterministic, the right hand side of (32) follows by multiplying and dividing the left hand side by the constant $F_0^{T_i} / F_0^P$.

$$E\left((S_{T_i} - K)^+\right) = \frac{F_0^{T_i}}{F_0^P} E\left(\left(F_{T_i}^P - K\frac{F_0^P}{F_0^{T_i}}\right)^+\right) \tag{32}$$

This is equivalent to consider that the implied volatilities of options on the spot S_t with strike *K* are the same as the implied volatilities on options on the forward F_t^P with an adjusted strike equal to $K F_0^P / F_0^{T_i}$ (note that these strikes change for each maturity). Table 2 shows the forward values $F_0^{T_i}$ valued at present time and delivered at each maturity T_i. From this table $F_0^P = 4107.9$ €.

TABLE 2: UNDERLYING FORWARD AT EACH MATURITY.

1m	**3m**	**6m**	**9m**	**1y**	**2y**	**3y**	**4y**	**5y**	**10y**
3870.6	3874.4	3880.3	3886	3892	3915.3	3938.9	3962.6	3986.5	4107.9

The bootstrapping algorithm of section 4 is implemented using the function "fminsearch" of the scientific package Matlab. The first calibration step searches for the parameters of the first period and the initial variance v_0. The initial variance is fixed at this first step for the rest of the calibrations. The weights of equation (30) are set to 100 for ATM options and 45, 35 and 5 as the moneyness gets apart the ATM. Call options are used for strikes greater than F_0^P and put options for strikes below. It has been observed that very different sets of parameters can fit the same market prices. This fact is not surprising as for a given volatility of variance, a sufficiently big mean reversion can produce the same result as a low mean reversion with very low volatility of variance. Therefore, the parameter seach space is limited to a set of intervals defined by the user so that more sensible parameters get out of the calibration. Table 3 presents the two search spaces that are considered. The set of intervals on the left represents a constrained search (especially with respect to σ and κ), whereas the set on the right represents an uncontrained search.

TABLE 3: SEARCH SPACE: CONSTRAINED (LEFT), UNCONSTRAINED (RIGHT).

	v_0	θ	κ	σ	ρ	v_0	θ	κ	σ	ρ
max	1	1	20	1.5	1	100	100	100	100	1
min	0	0	0	0	-1	0	0	0	0	-1

To avoid constrained optimisation, equation (33) shows a change of variable for which the constrained parameter p is expressed in terms of an unconstrained $\widetilde{p}$ for which the search is carried out. The constants $p_{\min}$ and $p_{\max}$ are the limits of the interval in which the parameter p is confined when $\widetilde{p}$ moves in the real line. The constant m has been set to 100 to make the transition of the hyperbolic tangent from –1 to 1 less abrupt.

$$p = \frac{p_{\max} - p_{\min}}{2}\left(1 + p_{\min} + \tanh\left(\frac{\widetilde{p}}{m}\right)\right) \qquad (33)$$

TABLE 4: CALIBRATED HESTON PARAMETERS: CONSTRAINED (UP), UNCONSTRAINED (DOWN).

P \ Mat	**1m**	**3m**	**6m**	**9m**	**1y**	**2y**	**3y**	**4y**	**5y**	**10y**
θ	0.01	0.03	0.03	0.03	0.05	0.05	0.07	0.12	0.14	0.31
κ	0.61	7.33	6.25	6.46	4.20	2.78	1.97	0.84	0.61	0.29
σ	0.60	0.56	1.13	1.15	1.09	1.26	1.18	1.14	1.12	1.14
ρ	-0.42	-0.46	-0.59	-0.63	-0.90	-0.67	-0.75	-0.77	-0.79	-0.84
θ	0.01	0.03	0.03	0.03	0.05	0.05	0.06	0.09	0.11	0.21
κ	0.84	4.75	3.08	5.21	4.87	4.82	3.81	3.89	4.51	3.02
σ	0.61	0.35	0.77	0.83	1.54	1.58	1.88	3.35	5.24	6.70
ρ	-0.42	-0.57	-0.56	-0.68	-0.77	-0.78	-0.80	-0.85	-0.88	-0.92

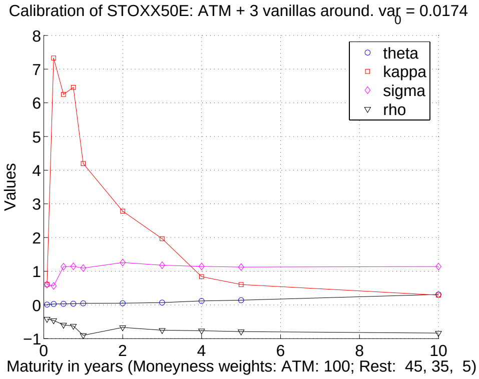


Fig. 3: Calibrated parameters with ***constrained*** *search space.*

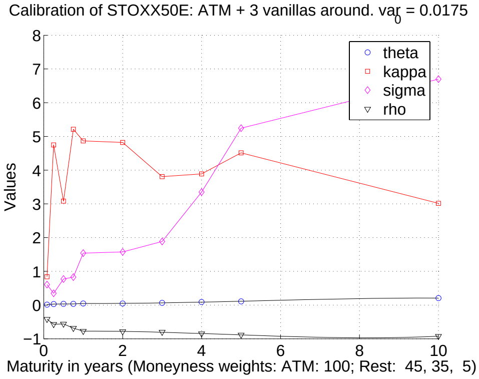


Fig. 4: Calibrated parameters with ***unconstrained*** *search space.*

Table 4, Fig. 3 and Fig. 4 present the calibrated parameters for the constrained $v_0 = 0.0174$ and the unconstrained $v_0 = 0.0175$ cases. The undiscounted vanilla option prices on F_t^P with spot equal to F_0^P and adjusted strike prices $K\,F_0^P / F_0^{T_i}$ for each moneyness and maturity are presented in Table 5. These prices are normalized by F_0^P, and expressed in basis points.

Table 6 shows the calibration error (market minus Heston) for each option in basis points for the constrained (above) and the unconstrained (below) cases. Note that these errors are all below 4 basis points except for the most out of the money options at long maturities. Both calibrations seem reasonable.

The interpretation of the time evolution of the parameters in terms of market expectations is tricky. Both calibrations suggest that the market is pricing in increasing volatility (increasing θ from short term volatility levels around 11% to 45% for 10 year maturity) and increasing skew (ρ progressively getting closer to −1). The unconstrained calibration suggests increasing uncertainty for the volatility (increasing σ) as the mean reversion is rather stable around 4. The constrained calibration forces the volatility of variance σ to be rather stable (the maximum level is 1.5) but the mean reversion κ progressively decreases indicating more long term uncertainty for the volatility. Therefore, from a qualitative point of view, both calibrations seem to agree that market is pricing in increasing volatility, increasing uncertainty for the volatility and increasing skew (more probability for outcomes with lower underlying levels).

TABLE 5: VANILLA OPTION PRICES (BASIS POINTS).

K \ Mat	1m	3m	6m	9m	1y	2y	3y	4y	5y	10y
0.85	0.4	12.8	60.1	115	181	412	632	850	1045	1958
0.90	3.7	36.8	115	190	271	537	780	1011	1216	2166
0.95	25.3	100	213	308	401	694	953	1194	1405	2384
1.00	138	255	383	488	584	887	1155	1399	1614	2612
1.05	11.7	79.7	189	295	396	747	1074	1378	1653	2851
1.10	0.7	18.5	72.2	144	220	532	846	1145	1418	2742
1.15	0.0	4.3	25.5	63.7	111	362	652	938	1205	2532

TABLE 6: CALIBRATION ERROR (BASIS POINTS): CONSTRAINED (UP), UNCONTRAINED (DOWN).

K \ Mat	1m	3m	6m	9m	1y	2y	3y	4y	5y	10y
0.85	1	1	-2	0	1	-4	0	-1	-3	-4
0.90	2	1	-1	-1	0	0	0	1	0	-1
0.95	-1	-1	1	0	0	1	0	0	0	-1
1.00	0	0	-1	0	0	1	0	0	1	2
1.05	0	0	0	0	-1	-1	0	0	-2	1
1.10	0	0	1	0	0	0	0	0	0	-2
1.15	0	0	0	-2	4	3	-1	1	4	-8
0.85	1	1	-1	1	0	-1	2	0	-1	-3
0.90	2	1	0	0	0	1	1	1	1	-1
0.95	-1	-1	1	0	0	1	0	0	0	-1
1.00	0	0	-1	0	0	0	0	-1	0	1
1.05	0	0	0	0	-1	-2	0	-1	-3	1
1.10	0	-1	0	0	0	1	1	1	1	-1
1.15	0	0	-1	-2	3	7	2	4	6	-7

The correct implementation of a Monte Carlo method for valuation of exotic products would require the use of an exact method such as [Broadie et al, 2004]. A regular Monte Carlo implemented with the Euler or even the Milstein method would not correctly work as the Feller condition ($2\kappa\theta > \sigma^2$) is not satisfied. This condition ensures that the variance process cannot reach zero. When the variance process reaches zero, an absorbing boundary condition is imposed. The discretization of the Monte Carlo cannot properly mimic this continuous absorbing condition and options are considerably overpriced. This bias increases with maturity and is not significantly reduced when the simulation time step shrinks. For the constrained case this bias is around 15% for 10 year maturity and 3% for 1 year maturity with 50 thousand simulations and a time step of 0.1 days. The bias error for the unconstrained case is around 100% for the 10 year maturity and 8% for the 1 year maturity. The error explodes at long maturities because σ is a lot higher. From a practical point of view, a constrained calibration that satisfied the Feller condition would be preferred, as conventional Monte Carlo methods would work. However with very skewed market scenarios (as the one analyzed here) this may not always be possible.

6 Application to forward start options

This section applies the methodology of sections 2 and 3 for the valuation of forward start vanilla options when the underlying follows Heston's process. This problem has already been addressed by [Lucic, 2003], solving a partial differencial equation similar to (50). The results are equivalent to those presented here. However, the approach of this section is straightforward and can be easily generalized to any model whose evolution can be expressed analitically in terms of a characteristic function of the form (2). Consider the forward start option of equation (34) which fixes the strike at time t_u and expires at time t_v according to Fig. 1.

$$p = P(0,t_v)E\left((e^{x_v} - Ke^{x_u})^+\right) = P(0,t_v)E\left(e^{x_u}\right)E\left((e^{\tilde{x}_v} - K)^+\right) \tag{34}$$

Consider the processes $\tilde{x}_t$ given by equation (35):

$$\tilde{x}_t = \begin{cases} 0 & t < u \\ x_t - x_u & t \geq u \end{cases} \tag{35}$$

The logarithmic evolution of the underlying up to time t_v $x_v = x_u + \tilde{x}_v$ and the variance process are decomposed in two parts: up to time t_u (x_t, v_t) and from t_u to t_v $(\tilde{x}_t, v_t)$. The first part is independent of the second only for the logarithm of the underlying but not for the variance. Therefore, e^{x_u} can be taken out of the expectation. The process $\tilde{x}_t$ is equal to cero at t_u but the variance process at t_u is unknown because it continues from the previous period. The distribution of $\tilde{\mathbf{x}}_t = (\tilde{x}_t, v_t)$ (for $t > u$) given $\tilde{\mathbf{x}}_u = (0, v_u)$ is known and defined by the characteristic funtion $\varphi_{uv}(\tilde{\mathbf{X}}/0, v_u)$ of equation (21), where $\tilde{\mathbf{X}} = (\tilde{X}, \tilde{V})$ and $\mathbf{x}_u = (0, v_u)$. The problem now reduces to find the distribution of $\tilde{\mathbf{x}}_t$ given the initial known state $\mathbf{x}_0 = (x_0, v_0)$.

Equation (36) shows that the density function $\tilde{f}(\tilde{\mathbf{x}}_t/0, v_0)$ of $\tilde{x}_t$ given v_0 can be calculated multiplying the density $f_{uv}(\tilde{\mathbf{x}}_t/0, v_u)$ of $\tilde{\mathbf{x}}_t$ given v_u by the probability of v_u $(f_{0u}(v_u/\mathbf{x}_0))$ and integrating over all possible v_u. Equation (37) shows the characteristic function to search.

$$\tilde{f}(\tilde{\mathbf{x}}_t/0, v_0) = \int_{\mathbf{R}^+} dv_u f_{0u}(v_u/\mathbf{x}_0) f_{uv}(\tilde{\mathbf{x}}_t/0, v_u) \tag{36}$$

$$\tilde{\varphi}(\tilde{\mathbf{X}}/0, v_0) = \int_{\mathbf{R}^2} e^{i\tilde{\mathbf{x}}_t \cdot \tilde{\mathbf{X}}} \tilde{f}(\tilde{\mathbf{x}}_t/0, v_0) d\tilde{\mathbf{x}}_t \tag{37}$$

If equation (36) is substituted in (37) and the sum order is exchanged (summing over $\tilde{\mathbf{x}}$ first), it yields equation (38).

$$\widetilde{\varphi}(\widetilde{\mathbf{X}}/0,\nu_0) = \int_{\mathbf{R}^+} d\nu_u f_{0u}(\nu_u/\mathbf{x}_0) \int_{\mathbf{R}^2} e^{i\widetilde{\mathbf{x}}_t\cdot\widetilde{\mathbf{X}}} f_{uv}(\widetilde{\mathbf{x}}_t/0,\nu_u)d\widetilde{\mathbf{x}}_t \tag{38}$$

The second integral is the definition of $\varphi_{uv}(\widetilde{\mathbf{X}}/0,\nu_u)$. Substituting this definition yields equation (39). Now the integral stands for the definition of the marginal characteristic function $\varphi_{0u}(V/\mathbf{x}_0)$ of the variance up to time t_u given by equation (23), setting X equal to 0. Substituting this definition yields equations (40) and the final characteristic function (41),

$$\widetilde{\varphi}(\widetilde{\mathbf{X}}/0,\nu_0) = e^{C_{uv}(\widetilde{\mathbf{X}})} \int_{\mathbf{R}^+} d\nu_u f_{0u}(\nu_u/\mathbf{x}_0)\, e^{i\left(-iD_{uv,2}(\widetilde{\mathbf{X}})\right)\nu_u} \tag{39}$$

$$\widetilde{\varphi}\left(\widetilde{X},\widetilde{V}/0,\nu_0\right) = e^{C_{uv}(\widetilde{X},\widetilde{V})}\varphi_{0u}\left(0,-iD_{uv,2}(\widetilde{X},\widetilde{V})/\mathbf{x}_0\right) \tag{40}$$

$$\widetilde{\varphi}\left(\widetilde{X},\widetilde{V}/0,\nu_0\right) = \exp\left(\widetilde{C}(\widetilde{X},\widetilde{V}) + \widetilde{D}(\widetilde{X},\widetilde{V})\nu_0\right) \tag{41}$$

where $\widetilde{C}$ and $\widetilde{D}$ are given by equation (42).

$$\begin{cases} \widetilde{C}\left(\widetilde{X},\widetilde{V}\right) = C_{uv}\left(\widetilde{X},\widetilde{V}\right) + C_{0u}\left(0,-iD_{uv,2}(\widetilde{X},\widetilde{V})\right) \\ \widetilde{D}\left(\widetilde{X},\widetilde{V}\right) = D_{0u,2}\left(0,-iD_{uv,2}(\widetilde{X},\widetilde{V})\right) \end{cases} \tag{42}$$

The marginal characteristic function of $\widetilde{x}_v$ is obtained setting $\widetilde{V}$ equal to zero and the price of the forward start option can be easily calculated using the same procedure of section 3 for vanilla options.

7 Forward Skew of Heston's Model

This section uses the results from section 6 to study the forward skew of both calibrations presented in section 5. By forward skew, it is understood the implied volatility surface that results from the forward start option price of equation (34). The implied volatility is the constant volatility σ_{BS} of the process (43) that used in (34) gives the same price as the forward start option when $\widetilde{x}_t$ is a Heston process.

$$\widetilde{x}_t = \left(\mu - \frac{1}{2}\sigma_{BS}{}^2\right)dt + \sigma_{BS}dW_t \tag{43}$$

As in section 5, the options considered are options on the forward F_t^P with spot equal to F_0^P and adjusted strike prices $KF_0^P/F_0^{T_i}$ for each moneyness and maturity. Prices are not discounted.

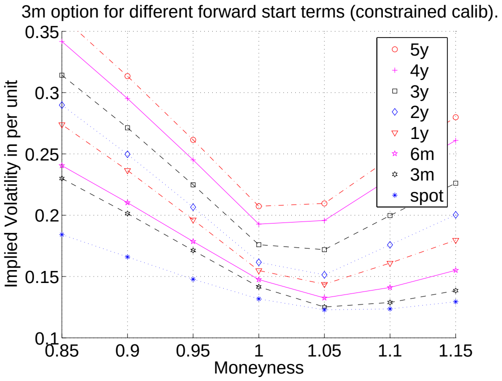


*Fig. 5: Implied volatility of a 3 month option for varying forward start terms and using the **constrained** calibration.*

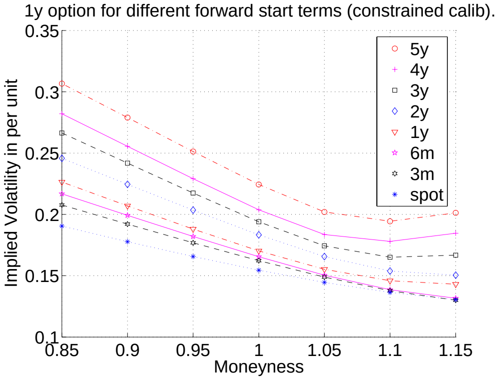


*Fig. 6: Implied volatility of a 1 year option for varying forward start terms and using the **constrained** calibration.*

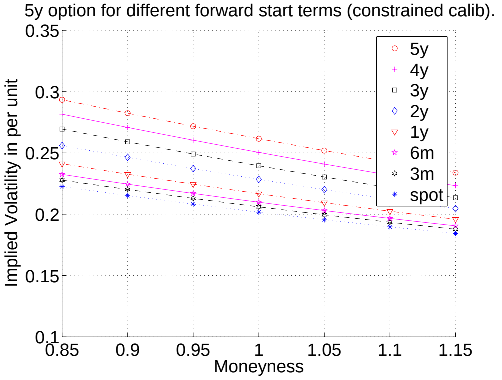


*Fig. 7: Implied volatility of a 5 year option for varying forward start terms and using the **constrained** calibration.*

Fig. 5 to Fig. 7 present the implied volatility of 3 month, 1 year and 5 year options when the forward start term changes and the constrained calibration of section 5 is used. The curve at the bottom corresponds to the spot option (the calibrated volatility from table 1). These figures show that the implied volatility and the slope of the skew increase with the forward term. This agrees well with the parameter interpretation of section 5 in which the market was pricing in increasing volatility, increasing uncertainty for the volatility and increasing skew.

Fig. 8 shows the implied volatility for the 3 month option using the unconstrained calibration. The implied volatility starts to decrease (especially around moneyness greater than 0.95) when the forward start term goes beyond 9 months. If Fig. 8 is compared with Fig. 5 (the same option valued with constrained calibration), the implied volatility for forward terms up to 9 months is very similar. This fact is confirmed by Fig. 9, which shows the price difference in basis points between the 3 month option valued with constrained and the same option valued with unconstrained calibration. For forward start terms up to 9 months, the price differences are below 10 basis points. However, very big differences appear when the forward start term goes beyond 9 months.

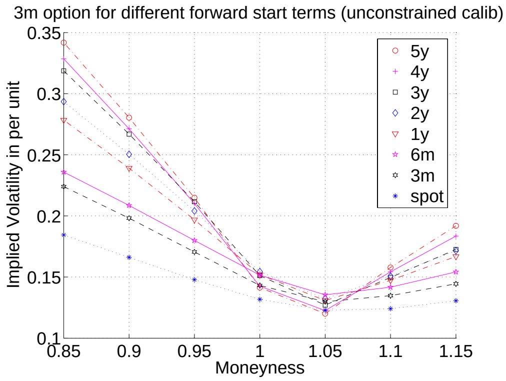


*Fig. 8: Implied volatility of a 3 month option for varying forward start terms and using the **unconstrained** calibration.*

The differences given by both calibrations are explained because the forward start options depend highly on the marginal distribution of the variance on the forward start date. This marginal distribution is not calibrated (only the marginal distribution of the underlying is calibrated). A long maturity option depends a lot less on the strike fixed at start. That is why the 5 year option of Fig. 10 valued with unconstrained parameters behaves much more alike Fig. 7 (the same option valued with constrained calibration).

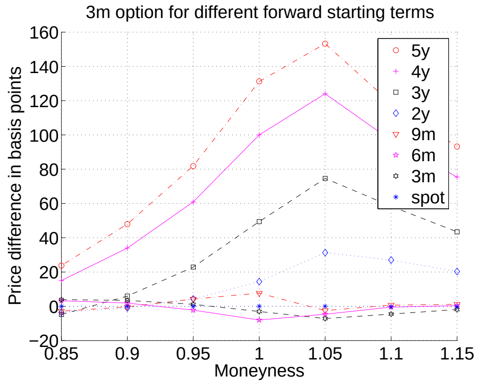


Fig. 9: Constrained minus unconstrained price in basis points of a 3 month option for varying forward start terms.

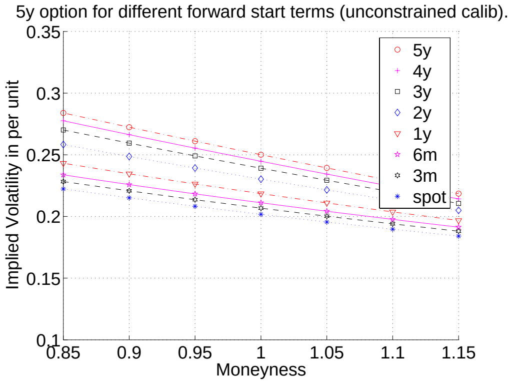


Fig. 10: Implied volatility of a 5 year option for varying forward start terms and using the ***unconstrained*** *calibration.*

The marginal distribution of the variance is a lot more skewed towards lower values for the unconstrained calibration. This fact justifies a lower total variance (the integral of the variance) and therefore lower prices and implied volatilities for the unconstrained calibration. To justify why the unconstrained calibration skews the variance towards lower values, consider the volatility process $\tilde{\sigma}_t$ in (44). This process is obtained applying Ito´s formula to the square root of Heston's variance process v_t in (18).

$$d\tilde{\sigma}_t = \left(\frac{4\kappa\theta - \sigma^2}{8\tilde{\sigma}_t} - \frac{\kappa}{2}\tilde{\sigma}_t \right) dt + \frac{1}{2}\sigma\, dY_t \tag{44}$$

The unconstrained calibration has greater κ and higher σ in comparison with the product $\kappa\theta$ for maturities beyond 9 months. This explains why the drift of $\tilde{\sigma}_t$ is considerably more negative for the unconstrained calibration (the drift is negative because the Feller condition is violated and therefore the zero variance point can be reached). See that when $\tilde{\sigma}_t$ equals zero the drift explodes (that´s why this equation cannot be used to integrate the variance process when the Feller condition is not satisfied).

If the forward start option were not so highly dependent on the distribution of the variance, both calibrations would give closer results (as those up to forward start terms of 9 months). However, it seems that the constrained calibration is considerably more reasonable. The results provided in this section show that it is indeed very important to keep in mind that the calibration is carried out only for the marginal distribution of the underlying (that is the only information provided by the market). Therefore, calibration should be carried out so that the forward skew makes sense to traders and practicioners using the model.

8 Conclusions

This paper presents a methodology to introduce piecewise constant time-dependent parameters preserving the analytic tractability for a wide family of models. This family includes hybrids with stochastic interest rates, stochastic volatility and jumps and their respective non-hybrid counterparts.

This methodology is built using transform methods based on analytic expressions of the characteristic function of the distribution of the underlying. The main contribution of the paper is the derivation of the characteristic function of the evolution of the underlying for a time horizon, in terms of the characteristic functions of the horizon sub-periods where the parameters change.

The method is applied to Heston's model to obtain a semi-analytical formula for valuation of vanilla options. A bootstrapping calibration algorithm is proposed and a case study works out the calibration of the volatility surface of the Eurostoxx 50 index.

The method is also applied to obtain a semi-analytical formula for valuation of forward start vanilla options driven by Heston's model. These formulas are used to explore the forward skew of the case study of the Eurostoxx 50 index.

Acknowledgement: the author of this paper wants to greatly thank the comments and suggestions of Eduard Giménez. They have improved considerably the organization and the materials presented in this paper. The author also wants to thank Eulogio Cuesta for his notation corrections.

References

[Albrecher et al, 2007] H. Albrecher, P. Mayer, W. Schoutens, J. Tistaert, "The Little Heston Trap", *Wilmott Magazine, Wiley & Sons*, January Issue, pp. 83-92.

[Andersen et al, 2002] L. B. G. Andersen, J. Andreasen, "Volatile volatilities", *Risk Magazine*, vol. 15, No. 12, pp. 163-168.

[Bakshi, 1997] G. Bakshi, C. Cao, Z. Chen, "Empirical performance of alternative option pricing models", *Journal of Finance*, vol. 52, No. 5, December 1997.

[Britten-Jones et al, 2000] M. Britten-Jones and A. Neuberger, “Option prices, implied price processes and stochastic volatility”, *Journal of Finance*, Vol. 55, no. 2, pp. 839-866, April 2000.

[Broadie et al, 2004] M. Broadie and O. Kaya, “Exact simulation of stochastic volatility and other affine jump diffusion processes, *Proceedings of the 2004 Winter Simulation Conference*, 2004.

[Carr et al, 1998] P. Carr, D. Madan, “Option valuation using the Fast Fourier Transform”, *Journal of Computational Finance*, vol. 2, pp. 61-73, 1998.

[Cox et al, 1985a] J. C. Cox, J. E. Ingersoll and S. A. Ross, “A Theory of the Term Structure of Interest Rates”, *Econometrica*, vol. 53, pp. 385-408, 1985.

[Davies, 1973] R. B. Davies, “Numerical Inversion of a Characteristic Function” *Biometrika*, vol. 60, pp. 415-417.

[Feller, 1971] W. Feller, “An Introduction to Probability Theory and its Applications”, *John Wiley and Sons*, vol. II, 2nd ed. New York, 1971.

[Heston, 1993] S. Heston, “A closed form solution for options with stochastic volatility with applications to bond and currency options”, *Review of Financial Studies*, No. 6, pp. 327-343, 1993.

[Kahl et al, 2005] C. Kahl, P. Jäckel, “Not so complex logarithms in the Heston model”, *Wilmott magazine*, September 2005, pp. 94-103.

[Kendal, 1987] M. Kendall, A. Stuart and J. K. Ord, “Kendall´s Advanced Theory of Statistics”, *Oxford University Press*, vol. 1, 5th ed., New York, 1987.

[Lee, 2005] R. W. Lee, “Option Pricing by Transform Methods: Extensions, Unification and Error Control”, *Journal of Computational Finance*, vol. 7, No. 3, pp. 51-86, 2004.

[Lord et al, 2006] R. Lord, C. Kahl, “Why the rotation count algorithm works”, Working Paper, University of Wuppertal, 2006.

[Lucic, 2003] V. Lucic, “Forward start options in stochastic volatility models”, *Wilmott magazine*, 2003.

[Merton, 1976] R. C. Merton, “Option Pricing when underlying stock returns are discontinuous¨, *Journal of Financial Economics*, No. 3, pp. 125-144, 1976.

[Mikhailov et al, 2005] S. Mikhailov, U. Nögel, “Heston´s Stochastic Volatility Model Implementation, Calibration and Some Extensions”, *Wilmott magazine*, 2003.

[Piterbarg, 2005] V. Piterbarg, “Stochastic Volatility Model with Time-dependent Skew”, *Journal of Applied Mathematical Finance*, Vol. 12, No 2, pp. 147-185, June 2005.

[Piterbarg, 2006] V. Piterbarg, “Time to Smile”, *Risk Magazine*, Vol. 19, No. 5, May 2006.

[Schoutens et al, 2004] W. Schoutens, E. Simons, J. Tistaert, “A perfect calibration! Now what?”, *Wilmott magazine*, March 2005, pp. 66-78.

[Scott, 1997] L. Scott, “Pricing stock options in a jump-diffusion model with stochastic volatility and interest rates: applications of Fourier inversion methods”, *Journal of Mathematical Finance*, vol. 7, No. 4, pp. 413-426, 1997.

[Shephard, 1991a] N. G. Shephard, “From Characteristic Function to Distribution Function: A Simple Framework for the Theory”, *Journal on Econometric Theory*, vol. 7, pp. 519-529, 1991.

[Shephard, 1991b] N. G. Shephard, “Numerical Integration Rules for Multivariate Inversions”, *Journal of Comput. Sim.*, vol. 39, pp. 37-46.

Appendix A: Deriving the characteristic function

Consider the process (45) in terms of x_t and ν_t with the same parameter definitions as (18).

$$\begin{cases} dx_t = \left(\mu - \frac{1}{2}\nu_t\right)dt + \sqrt{\nu_t}\,dW_t \\ d\nu_t = \kappa(\theta - \nu_t)dt + \sigma\sqrt{\nu_t}\,dY_t \end{cases} \qquad d\langle W_t, Y_t\rangle = \rho dt \tag{45}$$

The traditional way to calculate the expectation (46) is to integrate the payoff function g using a explicit formula for the density function of the joint probability distribution of x_T and ν_T given the initial values x_t and ν_t. Unfortunately, this density function is not analytic. However, [Heston, 1993] showed that it was possible to calculate the expectation h directly as the solution of a differential equation.

$$h(t, x_t, \nu_t) = \mathbf{E}\left(g(x_T, \nu_T)/x_t, \nu_t\right) \tag{46}$$

The characteristic function of the joint distribution would be given by the function h with payoff function (47).

$$g(X, V, x_T, \nu_T) = \exp\left(iXx_T + iV\nu_T\right) \tag{47}$$

In the appendix, [Heston, 1993] shows the derivation of the marginal characteristic function using the payoff function $g(X, x_T, \nu_T) = \exp(iXx_T)$. This appendix provides a more general solution in which the function g can provide payoffs not only of the marginal but also the joint characteristic function. The result presented here can also be found in [Mikhailov et al, 2005] but this paper only mentions that computer-algebra system Maple was used to obtain the result, but no derivation details are provided. Here, the whole derivation procedure is presented.

Consider the function h of equation (48), where $\mathbf{I}_t$ refers to the information set up to time t represented by the values of the stochastic process x and v at time t.

$$h(t, x_t, \nu_t) = \mathbf{E}\left(g(x_T, \nu_T)/x_t, \nu_t\right) = \mathbf{E}\left(g/\mathbf{I}_t\right) \tag{48}$$

Considering a time instant $s > t$ and applying the principle of iterated expectations, equation (49) shows that the function h is a martingale.

$$\mathbf{E}\left(h(s, x_s, \nu_s)/\mathbf{I}_t\right) = \mathbf{E}\left(\mathbf{E}(g/\mathbf{I}_s)/\mathbf{I}_t\right) = \mathbf{E}\left(g/\mathbf{I}_t\right) = h(t, x_t, \nu_t) \tag{49}$$

Applying Ito's lemma to h and forcing the drift to be zero (h is a martingale) gives the partial differential equation (50). This is indeed a very general result which can be applied to calculate the expectation of functions depending on any process.

$$\begin{aligned}&\frac{\partial h}{\partial t}+\frac{\partial h}{\partial x_t}\left(\mu-\frac{1}{2}\nu_t\right)+\frac{\partial h}{\partial \nu_t}\kappa(\theta-\nu_t)\\&\quad+\frac{1}{2}\frac{\partial^2 h}{\partial x_t^2}\nu_t+\frac{1}{2}\frac{\partial^2 h}{\partial \nu_t^2}\sigma^2\nu_t+\frac{\partial^2 h}{\partial x_t\partial \nu_t}\sigma\rho\nu_t=0\end{aligned}\qquad(50)$$

To determine the solution of equation (50), the final condition (51) at time *T* must be specified.

$$h(T,x_T,\nu_T)=g(x_T,\nu_T)\qquad(51)$$

The final payoff function that will be considered has the form (52), where three additional parameters have been introduced: *X*, C^0 and D^0. As already mentioned in equation (47), if $C^0=0$ and $D^0=iV$, the resulting payoff corresponds to the characteristic function of the joint distribution.

$$g(X,C^0,D^0,x_T,\nu_T)=\exp\left(C^0+D^0\nu_T+iXx_T\right)\qquad(52)$$

Equation (53) shows the solution of equation (50) guessed by [Heston, 1993],

$$h(t,x_T,\nu_T)=\exp\left(C+D\nu_T+iXx_T\right)\qquad(53)$$

where *C* and *D* are functions that depend, according to equation (54), on time to maturity $\tau=T-t$, *X*, C^0, D^0 and all the model parameters in (45), omitted here to simplify notation.

$$C=C(\tau,X,C^0,D^0)\quad D=D(\tau,X,D^0,C^0)\qquad(54)$$

Substituting the tentative solution (53) in (50) yields (55), where *A*, *B* and *M* are given by (56).

$$-\frac{\partial C}{\partial t}-iX\mu-\kappa\theta D+\left(-\frac{\partial D}{\partial t}+AD^2+BD+M\right)\nu_t=0\qquad(55)$$

$$A=-\frac{1}{2}\sigma^2\quad B=\kappa-iX\sigma\rho\quad M=\frac{1}{2}X(i+X)\qquad(56)$$

As ν_t is stochastic, expression (55) will be zero only if both the term multiplying ν_t and the other one are zero independently. On the other hand it is more convenient to use $\tau=T-t$ as parameter rather than *t*. Therefore, the negative of the partial derivatives with respect to τ will replace the partial derivatives with respect to *t*. This leads to the system of differential equations (57) and (58) (for the purpose of solving it *C* and *D* are functions of τ). The terminal condition for this system of equations is given by $C^0=C(\tau=0)$ and $D^0=D(\tau=0)$ so that $h(T,x_T,\nu_T)$ becomes the final payoff function (52).

$$\frac{\partial D}{\partial \tau}+AD^2+BD+M=0\qquad(57)$$

$$\frac{\partial C}{\partial \tau}-iX\mu-\kappa\theta D=0\qquad(58)$$

Expression (57) is a Riccati equation that only depends on *D*. This Riccati equation can be turned into the ordinal differential equation (59) through the change of variable $z=(D-D_1)^{-1}$, where D_1 is a particular solution.

$$\frac{\partial z}{\partial \tau}+Lz-A \quad \text{with} \quad L=-(2AD_1+B) \tag{59}$$

The solution of the ordinal equation (59) is given by (60):

$$z=U+W\exp(-L\tau) \quad U=\frac{A}{L} \quad W=\left(\frac{1}{D^0-D_1}-\frac{A}{L}\right) \tag{60}$$

Undoing the change of variables yields the solution (61).

$$D=\frac{1+D_1(U+W\exp(-L\tau))}{U+W\exp(-L\tau)} \tag{61}$$

Nothing has been said about the particular solution D_1 yet. Indeed, if constant solutions were considered, D_1 would be the solution of the second order equation (62) after substituting it in equation (57).

$$AD_1^2+BD_1+M=0 \tag{62}$$

This equation has two solutions: taking the positive square root of the second order equation and substituting in (59) yields $L=-d$, where *d* is given by equation (64). Taking the negative square root yields $L=d$. The solution used by [Heston, 1993] is $L=-d$. [Albrecher et al, 2007] presents an extensive study proving that both solutions are completely equivalent from a theoretical point of view. However, using $L=-d$ gives plenty of numerical problems (especially for long maturities) as reported in [Kahl et al, 2005], whereas the second solution where $L=d$ avoids them all (see [Albrecher et al, 2007] for a rigorous proof; [Lord et al, 2006] reach the same solution using a different technique under certain parameter restrictions). An intuitive way of realizing that $L=d$ is a better choice is because the exponentials in (61) are decaying. This means that the complex exponential will not oscillate as the maturity increases and the modulus would not explode at long maturities. After simple but tedious algebraic manipulations and choosing $L=d$, equation (61) turns into the final solution (63) where the unknown parameters are given in (64). Please, note that if this expression is compared with that of [Heston, 1993], *d* will appear with the sign changed (it is not a misprint). In addition, C^0 and D^0 come out, generalizing the result to allow for more flexible payoffs.

$$D=\frac{\kappa-\rho\sigma Xi+d}{\sigma^2}\left(\frac{g-\tilde{g}e^{-d\tau}}{1-\tilde{g}e^{-d\tau}}\right) \tag{63}$$

$$\begin{gathered}\tilde{g}=\frac{\kappa-\rho\sigma Xi-d-D^0\sigma^2}{\kappa-\rho\sigma Xi+d-D^0\sigma^2} \qquad g=\frac{\kappa-\rho\sigma Xi-d}{\kappa-\rho\sigma Xi+d}\\ d=\sqrt{(\kappa-\rho\sigma Xi)^2+\sigma^2X(i+X)}\end{gathered} \tag{64}$$

If the solution (63) is substituted in equation (58), the resulting equation is the ordinal differential equation (65), where α and β are given in (66).

$$\frac{\partial C}{\partial \tau} - \alpha \frac{g - \tilde{g} e^{-d\tau}}{1 - \tilde{g} e^{-d\tau}} - \beta = 0 \tag{65}$$

$$\alpha = \kappa\theta \frac{\kappa - \rho\sigma X i + d}{\sigma^2} \qquad \beta = iX\mu \tag{66}$$

The solution is given by equation (67), where K_C is a constant that will be calculated to satisfy the terminal condition $C(\tau = 0) = C^0$.

$$C = \int \alpha \frac{g - \tilde{g} e^{-d\tau}}{1 - \tilde{g} e^{-d\tau}} \partial\tau + \beta\tau + K_C = \alpha I + \beta\tau + K_C \tag{67}$$

The indefinite integral I can be calculated doing the change of variable $u = \exp(-d\tau)$ and expanding the result in partial fractions with known integral (a logarithm). Equation (68) shows the final result.

$$\int \frac{g - \tilde{g} e^{-d\tau}}{1 - \tilde{g} e^{-d\tau}} \partial\tau = \frac{-1}{d} \int \left(\frac{(g-1)\tilde{g}}{1 - \tilde{g} u} + \frac{g}{u} \right) du = \frac{g-1}{d} \ln(1 - \tilde{g} e^{-d\tau}) + g d\tau \tag{68}$$

Equation (69) shows the constant from imposing the terminal condition $C(\tau = 0) = C^0$.

$$K_C = C^0 - \alpha \frac{g-1}{d} \ln\left(1 - \tilde{g}\right) \tag{69}$$

Replacing (68) and (69) in (67) gives the final result (70).

$$C = i\mu X\tau + \frac{\kappa\theta}{\sigma^2} \left(\ln\left(\frac{1 - \tilde{g}\, e^{-d\tau}}{1 - \tilde{g}} \right)^{-2} + \left(\kappa - \rho\sigma X i - d\right)\tau \right) + C^0 \tag{70}$$

If this result is compared with that of [Heston, 1993], C^0 and $\tilde{g}$ come up (they incorporate the initial conditions C^0 and D^0). In addition *d* appears with the sign changed (it is not a misprint) as already discussed for equation (63).

A. Elices obtained a PhD in Power Systems engineering at Pontificia Comillas University (Madrid, Spain) and a Masters in Financial Mathematics in the University of Chicago. He is a senior quant team member of the model validation group of the Risk Department at Grupo Santander in Madrid after working in a hedge fund in New York.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Numerical analysis of a quasistatic elasto-piezoelectric contact problem with damage

José R. Fernández[1] **and Rebeca Martínez**[1]

[1] *Departamento de Matemática Aplicada, Universidade de Santiago de Compostela*

emails: `jramon@usc.es`, `rebeca.martinez2@rai.usc.es`

Abstract

In this work, a contact problem between an elasto-piezoelectric body and a deformable obstacle is numerically studied. The damage of the material, caused by internal tension or compression, is also included into the model. The variational formulation leads to a coupled system composed of a nonlinear variational equation for the displacement field, a linear variational equation for the electric potential, and a nonlinear parabolic variational equation for the damage field. The existence of a unique weak solution is recalled. Then, a fully discrete scheme is introduced by using a finite element method to approximate the spatial variable and an Euler scheme to discretize the time derivatives. Error estimates are derived on the approximate solutions, and the linear convergence of the algorithm is derived under suitable regularity conditions. Finally, a two-dimensional example is presented to demonstrate the behaviour of the solution.

Key words: Elasto-piezoelectricity, damage, normal compliance, finite elements, error estimates, numerical simulations.
MSC 2000: 74B20, 74R05, 74M15, 65M15, 65M60, 74S05

Introduction

In this paper, we study, from the numerical point of view, the mechanical deformation of an elasto-piezoelectric body, taking into account the material damage into the model.

This work has three main aspects. First, we consider the effect of the damage, due to mechanical stress or strain, which appears in many engineering applications where the forces acting on the system vary periodically, and it leads to the decrease in the load carrying capacity of the body. There exists very large engineering literature dealing with the many approaches and facets of material damage (see, e.g., [17, 18, 20]). Other models for mechanical damage derived from thermomechanical considerations appeared in [9, 10]. In this approach, the damage field ζ varies between one and zero at each point of the body, in such a way that when $\zeta = 1$ the material is damage-free, when $\zeta = 0$ the material is completely damaged and when $0 < \zeta < 1$ there is partial damage.

Secondly, piezoelectricity is also considered. It is usually defined as the ability of certain crystals, like quartz or even some human bones, to produce a voltage when they are subjected to mechanical stress. The piezoelectric effect is characterized by the coupling between the mechanical and the electrical properties of the material: it was observed that the appearance of electric charges on some crystals was due to the action of body forces and surface tractions (this is called the *sensor effect*). Conversely, the action of the electric field generated strain or stress in the body (the so-called *actuator effect*). Different models have been developed to describe the interaction between the electric and mechanical fields (see, e.g., [2, 14, 21, 22] and the references therein).

Recently, some attempts have been made to study the coupling between the damage and the piezoelectric effects. For instance, a model was proposed in [23] by assuming that the damage affects both the elastic tensor and the electric displacements. Here, we use this model and we continue the investigation reported in [6, 7], where the contact was not considered. This is assumed with a deformable obstacle and it constitutes the third aspect of this work. According to [15, 19], the well-known normal compliance contact condition was employed for its modelling. The numerical analysis of the corresponding variational problem is provided in this paper, and we also perform some two-dimensional numerical simulations to show the piezoelectric behaviour.

The paper is structured as follows. In Section 1 we briefly describe the model for the process and we present its variational formulation. Then, in Section 2 we provide the numerical analysis of the weak problem, and we prove a main error estimates result, Theorem 2, from which the linear convergence of the fully discrete scheme is deduced under suitable regularity conditions. Finally, in Section 3 some numerical results, involving test examples in two dimensions, are shown to demonstrate the behaviour of the model.

1 Mechanical problem and variational formulation

In this section, we present a brief description of the mechanical problem (full details concerning the model can be found in [6], and we refer the reader to [8] as an example of the normal compliance contact condition).

Let $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ be a domain occupied by an elasto-piezoelectric body with outer surface $\partial\Omega = \Gamma$, assumed to be sufficiently smooth and decomposed into three disjoint measurable parts Γ_D, Γ_N and Γ_C, such that meas $(\Gamma_D) > 0$. For each $\boldsymbol{x} \in \Gamma$, let $\boldsymbol{\nu}(\boldsymbol{x})$ be the unit normal outward vector to Γ. Let us denote by $[0, T]$, $T > 0$, the time interval of interest. Volume forces of density $\boldsymbol{f}_B$ act in $\Omega \times (0, T)$ and volume electric charges of density q_B are present in $\Omega \times (0, T)$. Traction forces of density $\boldsymbol{f}_N$ act on $\Gamma_N \times (0, T)$ and surface electric charges of density q_N are found on $\Gamma_N \times (0, T)$. Finally, we assume that the body may come in contact with a deformable insulator obstacle on the boundary part Γ_C which is located at a distance s, measured along the outward unit normal vector $\boldsymbol{\nu}$ (see Figure 1).

Let $\boldsymbol{x} \in \Omega$ and $t \in [0, T]$ be the spatial and time variables, respectively. In order to simplify the writing, we do not indicate the dependence of the functions on $\boldsymbol{x}$ and

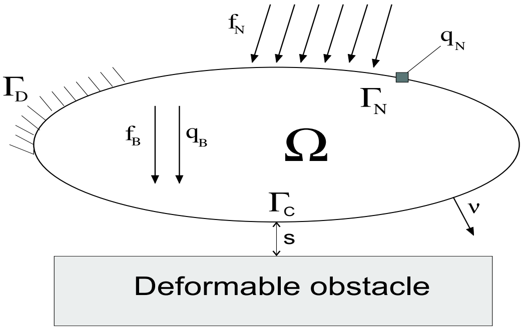


Figure 1: A damageable elasto-piezoelectric body in contact with a deformable obstacle.

t. Moreover, a prime after a variable represents the derivative with respect to the time variable.

We denote the displacement field, the stress tensor, the linearized strain tensor and the electric potential by $\boldsymbol{u}$, $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ and φ, respectively. We let ζ denote the damage field, which is defined in $\Omega\times(0,T)$ and measures the fractional decrease in the strength of the material. The material is assumed elasto-piezoelectric with constitutive law (see [5, 23]),

$$\boldsymbol{\sigma}=\zeta\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u})-\zeta\mathcal{E}^*\mathbf{E}(\varphi),$$

where $\mathcal{A}$ is the fourth-order elasticity tensor and $\mathbf{E}(\varphi)=(E_i(\varphi))_{i=1}^d$ represents the electric field defined by

$$E_i(\varphi)=-\frac{\partial\varphi}{\partial x_i},\quad i=1,\dots,d,$$

and $\mathcal{E}^*=(e^*_{ijk})_{i,j,k=1}^d$ denotes the transpose of the third-order piezoelectric tensor $\mathcal{E}=(e_{ijk})_{i,j,k=1}^d$. We recall that

$$e^*_{ijk}=e_{kij},\quad\text{for all}\quad i,j,k=1,\dots,d,$$

and the classical linearized elasto-piezoelectricity is obtained when $\zeta\equiv1$.

According to [2, 23], the following constitutive law is employed for the electric potential,

$$\mathbf{D}=\zeta\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u})+\zeta\beta\mathbf{E}(\varphi),$$

where $\mathbf{D}$ and β are the electric displacement field and the electric permittivity tensor, respectively.

We now describe the damage process. As a result of the tensile or compressive stresses in the body, micro-cracks and micro-cavities open and grow and this causes the load bearing capacity of the material to decrease. This reduction in the strength of an isotropic material is modelled by introducing the damage field $\zeta=\zeta(\boldsymbol{x},t)$ as the ratio

$$\zeta=\zeta(\boldsymbol{x},t)=\frac{E_{eff}}{E}$$

between the effective Young's modulus of elasticity E_{eff} and that of the damage-free material E. Obviously, it follows that $0 \leq \zeta \leq 1$.

Following the derivation presented in [9, 10], the evolution of the microscopic cracks and cavities responsible for the damage is described by the following parabolic partial differential equation,

$$\zeta' - \kappa \triangle \zeta = \phi(\boldsymbol{\varepsilon}(\boldsymbol{u}), \zeta).$$

Here, $\triangle$ denotes the Laplace operator, $\kappa > 0$ is the damage diffusion constant and ϕ represents the damage source function. Moreover, we assume that there is no damage influx throughout the boundary Γ and therefore, $\partial\zeta/\partial\boldsymbol{\nu} = 0$ on Γ.

Next, we describe the boundary conditions for the displacements and the electric potential field.

On the boundary part Γ_D we assume that the body is clamped (that is, $\boldsymbol{u} = \boldsymbol{0}$ on $\Gamma_D \times (0,T)$), and that it is subjected to a prescribed electric potential φ_D. A density of traction forces, denoted by $\boldsymbol{f}_N$, acts on the boundary part Γ_N and so,

$$\boldsymbol{\sigma\nu} = \boldsymbol{f}_N \quad \text{on} \quad \Gamma_N \times (0,T).$$

Moreover, we assume that surface electric charges of density q_N are applied on Γ_N; that is,

$$\mathbf{D} \cdot \boldsymbol{\nu} = q_N \quad \text{on} \quad \Gamma_N \times (0,T).$$

Finally, according to [15, 19], since the contact is assumed with a deformable obstacle, the following normal compliance contact condition is employed,

$$-\sigma_\nu = p(u_\nu - s),$$

where p is the normal compliance function whose properties will be described below, the normal stress is given by $\sigma_\nu = \boldsymbol{\sigma\nu} \cdot \boldsymbol{\nu}$ and $u_\nu = \boldsymbol{u} \cdot \boldsymbol{\nu}$ denotes the normal displacement in such a way that, when $u_\nu > s$, the difference $u_\nu - s$ represents the interpenetration of the body's asperities into those of the obstacle. We also assume that the contact is frictionless, i.e. the tangential component of the stress field, denoted $\boldsymbol{\sigma}_\tau = \boldsymbol{\sigma\nu} - \sigma_\nu\boldsymbol{\nu}$, vanishes on this contact surface.

For technical reasons associated with the loss of coercivity in the elastic equation, and possible singularities in ϕ as $\zeta \to 0$, we introduce the truncation operator η_*. This is a nondecreasing function which has the following form, for a fixed $\zeta_* > 0$,

$$\eta_*(\zeta) = \begin{cases} 1 & \text{if} \quad \zeta > 1, \\ \zeta & \text{if} \quad \zeta_* \leq \zeta \leq 1, \\ \zeta_* & \text{if} \quad \zeta < \zeta_*. \end{cases}$$

We note that as long as $\zeta \in [\zeta_*, 1]$ it makes no difference whether we use ζ or $\eta_*(\zeta)$. The existence of such lower limit for the damage, ζ_*, is justified because, when the damage is large, a crack can be generated and the linear elasticity theory can not be applied.

The mechanical problem corresponding to the quasistatic damage evolution in an elasto-piezoelectric body in contact with a deformable obstacle is written as follows.

Problem *P*. Find a displacement field $\boldsymbol{u} : \Omega \times (0,T) \to \mathbb{R}^d$, a stress field $\boldsymbol{\sigma} : \Omega \times (0,T) \to \mathbb{S}^d$, an electric potential field $\varphi : \Omega \times (0,T) \to \mathbb{R}$, an electric displacement field $\mathbf{D} : \Omega \times (0,T) \to \mathbb{R}^d$ and a damage field $\zeta : \Omega \times (0,T) \to \mathbb{R}$ such that,

$$-\mathrm{Div}\,\boldsymbol{\sigma} = \boldsymbol{f}_B \quad \text{in} \quad \Omega \times (0,T), \tag{1}$$

$$\mathrm{div}\,\mathbf{D} = q_B \quad \text{in} \quad \Omega \times (0,T), \tag{2}$$

$$\boldsymbol{\sigma} = \eta_*(\zeta)\,\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}) - \eta_*(\zeta)\,\mathcal{E}^*\mathbf{E}(\varphi) \quad \text{in} \quad \Omega \times (0,T), \tag{3}$$

$$\mathbf{D} = \eta_*(\zeta)\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}) + \eta_*(\zeta)\beta\mathbf{E}(\varphi) \quad \text{in} \quad \Omega \times (0,T), \tag{4}$$

$$\zeta' - \kappa\Delta\zeta = \phi\left(\boldsymbol{\varepsilon}(\boldsymbol{u}), \eta_*(\zeta)\right) \quad \text{in} \quad \Omega \times (0,T), \tag{5}$$

$$\frac{\partial\zeta}{\partial\boldsymbol{\nu}} = 0 \quad \text{on} \quad \Gamma \times (0,T), \tag{6}$$

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on} \quad \Gamma_D \times (0,T), \tag{7}$$

$$\boldsymbol{\sigma\nu} = \boldsymbol{f}_N \quad \text{on} \quad \Gamma_N \times (0,T), \tag{8}$$

$$\varphi = \varphi_D \quad \text{on} \quad \Gamma_D \times (0,T), \tag{9}$$

$$\mathbf{D}\cdot\boldsymbol{\nu} = q_N \quad \text{on} \quad \Gamma_N \times (0,T), \tag{10}$$

$$\boldsymbol{\sigma}_\tau = \boldsymbol{0}, \quad -\sigma_\nu = p(u_\nu - s) \quad \text{on} \quad \Gamma_C \times (0,T), \tag{11}$$

$$\zeta(0) = \zeta_0 \quad \text{in} \quad \Omega. \tag{12}$$

Here, ζ_0 represents an initial condition for the damage field, $\mathbb{S}^d$ denotes the space of symmetric $d \times d$ matrices with the usual notation of inner product and Div and div denote the divergence operators for tensor or vector functions, respectively.

We now present the variational formulation of the problem. Let $Y = L^2(\Omega)$, $H = [L^2(\Omega)]^d$, and denote by Q the space of second order symmetric tensor functions,

$$Q = \left\{\boldsymbol{\tau} \in [L^2(\Omega)]^{d\times d}\,;\ \tau_{ij} = \tau_{ji}, \quad i,j = 1,\ldots,d\right\}.$$

Let V and W be the variational spaces defined by

$$V = \{\boldsymbol{w} \in [H^1(\Omega)]^d\,;\, \boldsymbol{w} = \boldsymbol{0} \quad \text{on} \quad \Gamma_D\},$$

$$W = \{\psi \in H^1(\Omega)\,;\, \psi = 0 \quad \text{on} \quad \Gamma_D\},$$

and denote by W_D the subset of $H^1(\Omega)$ given by

$$W_D = \{\psi \in H^1(\Omega)\,;\, \psi = \varphi_D \quad \text{on} \quad \Gamma_D\}.$$

We now describe the assumptions on the problem data.
The elasticity tensor $\mathcal{A} = (a_{ijkl}(\boldsymbol{x}))_{i,j,k,l=1}^d : \boldsymbol{\tau} \in \mathbb{S}^d \to \mathcal{A}(\boldsymbol{x})(\boldsymbol{\tau}) \in \mathbb{S}^d$ satisfies:

$$\left.\begin{array}{l}
\text{(a) } a_{ijkl} = a_{klij} = a_{jikl} \quad \text{for} \quad i,j,k,l = 1,\ldots,d.\\
\text{(b) } a_{ijkl} \in L^\infty(\Omega) \quad \text{for} \quad i,j,k,l = 1,\ldots,d.\\
\text{(c) There exists } m_{\mathcal{A}} > 0 \text{ such that } \mathcal{A}(\boldsymbol{x})\boldsymbol{\tau}\cdot\boldsymbol{\tau} \geq m_{\mathcal{A}}\,\|\boldsymbol{\tau}\|^2\\
\quad\ \forall\,\boldsymbol{\tau} \in \mathbb{S}^d,\ \text{a.e. } \boldsymbol{x} \in \Omega.
\end{array}\right. \tag{13}$$

The piezoelectric tensor $\mathcal{E}(\boldsymbol{x}) = (e_{ijk}(\boldsymbol{x}))_{i,j,k=1}^d : \boldsymbol{\tau} \in \mathbb{S}^d \to \mathcal{E}(\boldsymbol{x})(\boldsymbol{\tau}) \in \mathbb{R}^d$ satisfies:

$$\left.\begin{array}{l}
\text{(a) } e_{ijk} = e_{ikj} \quad \text{for} \quad i,j,k = 1,\ldots,d.\\
\text{(b) } e_{ijk} \in L^\infty(\Omega) \quad \text{for} \quad i,j,k = 1,\ldots,d.
\end{array}\right. \tag{14}$$

The permittivity tensor $\beta(\boldsymbol{x}) = (\beta_{ij}(\boldsymbol{x}))_{i,j=1}^{d} : \boldsymbol{w} \in \mathbb{R}^d \to \beta(\boldsymbol{x})(\boldsymbol{w}) \in \mathbb{R}^d$ satisfies:

$$\begin{array}{l} \text{(a) } \beta_{ij} = \beta_{ji} \quad \text{for} \quad i, j = 1, \dots, d. \\ \text{(b) } \beta_{ij} \in L^\infty(\Omega) \quad \text{for} \quad i, j = 1, \dots, d. \\ \text{(c) There exists } m_\beta > 0 \text{ such that } \beta(\boldsymbol{x})\boldsymbol{w} \cdot \boldsymbol{w} \geq m_\beta \|\boldsymbol{w}\|^2 \\ \qquad \forall\, \boldsymbol{w} \in \mathbb{R}^d, \text{ a.e. } \boldsymbol{x} \in \Omega. \end{array} \tag{15}$$

The normal compliance function $p(\boldsymbol{x}) : r \in \mathbb{R} \to p(\boldsymbol{x})(r) \in [0, \infty)$ satisfies:

$$\begin{array}{l} \text{(a) There exists } m_p > 0 \text{ such that} \\ \qquad |p(\boldsymbol{x}, r_1) - p(\boldsymbol{x}, r_2)| \leq m_p\, |r_1 - r_2| \\ \qquad\qquad \forall\, r_1, r_2 \in \mathbb{R}, \text{ a.e. } \boldsymbol{x} \in \Gamma_C. \\ \text{(b) } (p(\boldsymbol{x}, r_1) - p(\boldsymbol{x}, r_2))(r_1 - r_2) \geq 0 \quad \forall\, r_1, r_2 \in \mathbb{R}, \text{ a.e. } \boldsymbol{x} \in \Gamma_C. \\ \text{(c) The mapping } \boldsymbol{x} \in \Gamma_C \mapsto p(\boldsymbol{x}, r) \text{ is measurable on } \Gamma_C, \\ \qquad \text{for all } r \in \mathbb{R}. \\ \text{(d) } p(\boldsymbol{x}, r) = 0 \quad \text{for all } r \leq 0. \end{array} \tag{16}$$

The damage source function $\phi : \Omega \times \mathbb{S}^d \times \mathbb{R} \to \mathbb{R}$ satisfies:

$$\begin{array}{l} \text{(a) There exists } L_\phi > 0 \text{ such that} \\ \qquad |\phi\,(\boldsymbol{x}, \boldsymbol{\varepsilon}_1, \zeta_1) - \phi\,(\boldsymbol{x}, \boldsymbol{\varepsilon}_2, \zeta_2)| \leq L_\phi\,(|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2| + |\zeta_1 - \zeta_2|) \\ \qquad \text{for all } \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}^d, \ \ \zeta_1, \zeta_2 \in \mathbb{R}, \quad \text{a.e.} \ \ \boldsymbol{x} \in \Omega. \\ \text{(b) The function } \ \boldsymbol{x} \to \phi\,(\boldsymbol{x}, \boldsymbol{\varepsilon}, \ \zeta) \ \text{ is measurable.} \\ \text{(c) The mapping } \boldsymbol{x} \to \phi\,(\boldsymbol{x}, \boldsymbol{0}, \zeta_*) \ \text{ belongs to } Y. \\ \text{(d) } \phi\,(\boldsymbol{x}, \boldsymbol{\varepsilon}, \zeta) \ \text{ is bounded.} \\ \text{(e) } \phi\,(\boldsymbol{\varepsilon}, \zeta) \leq 0 \ \text{ if } \ \zeta \geq 1, \quad \phi\,(\boldsymbol{\varepsilon}, \zeta) \geq 0 \ \text{ if } \ \zeta \leq \zeta_*. \end{array} \tag{17}$$

The following regularity is assumed on the density of volume forces, tractions, volume electric charges and surface electric charges:

$$\begin{array}{ll} \boldsymbol{f}_B \in C([0, T]; H), & \boldsymbol{f}_N \in C([0, T]; [L^2(\Gamma_N)]^d), \\ q_B \in C([0, T]; Y), & q_N \in C([0, T]; L^2(\Gamma_N)). \end{array} \tag{18}$$

Using Riesz' Theorem, we define the linear mappings $\boldsymbol{f} : [0, T] \to V$ and $q : [0, T] \to W$ as follows,

$$(\boldsymbol{f}(t), \boldsymbol{w})_V = \int_\Omega \boldsymbol{f}_B(t) \cdot \boldsymbol{w}\, d\boldsymbol{x} + \int_{\Gamma_N} \boldsymbol{f}_N(t) \cdot \boldsymbol{w}\, d\Gamma \quad \forall \boldsymbol{w} \in V,$$
$$(q(t), \psi)_W = \int_\Omega q_B(t) \psi\, d\boldsymbol{x} - \int_{\Gamma_N} q_N(t) \psi\, d\Gamma \quad \forall \psi \in W.$$

Let us denote by $j : V \times V \to \mathbb{R}$ the normal compliance functional given by

$$j(\boldsymbol{u}, \boldsymbol{w}) = \int_{\Gamma_C} p(u_\nu - s)\, w_\nu\, d\Gamma \quad \forall \boldsymbol{u}, \boldsymbol{w} \in V,$$

where we use the notation $w_\nu = \boldsymbol{w} \cdot \boldsymbol{\nu}$ for all $\boldsymbol{w} \in V$.

Finally, we assume that the initial condition ζ_0 satisfies

$$\zeta_0 \in H^1(\Omega), \quad \zeta_0\,(\boldsymbol{x}) \in (\zeta_*, 1] \quad \text{a.e.} \quad \boldsymbol{x} \in \Omega, \tag{19}$$

and define the bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ given by

$$a(\xi, \eta) = \kappa \int_\Omega \nabla\xi \cdot \nabla\eta \, d\boldsymbol{x} \quad \forall \xi, \eta \in H^1(\Omega).$$

Using Green's formula and boundary conditions (6)-(11), the variational formulation of Problem P is then written as follows.

Problem VP. *Find a displacement field* $\boldsymbol{u} : [0,T] \to V$*, an electric potential field* $\varphi : [0,T] \to W_D$ *and a damage field* $\zeta : [0,T] \to H^1(\Omega)$ *such that* $\zeta(0) = \zeta_0$ *and for a.e.* $t \in (0,T)$*,*

$$(\eta_*(\zeta(t))[\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}(t)) + \mathcal{E}^*\nabla\varphi(t)], \boldsymbol{\varepsilon}(\boldsymbol{w}))_Q + j(\boldsymbol{u}(t), \boldsymbol{w}) = (\boldsymbol{f}(t), \boldsymbol{w})_V \quad \forall \boldsymbol{w} \in V, \tag{20}$$

$$(\zeta'(t), \xi)_Y + a(\zeta(t), \xi) = (\phi(\boldsymbol{\varepsilon}(\boldsymbol{u}(t)), \zeta(t)), \xi)_Y \quad \forall \xi \in H^1(\Omega), \tag{21}$$

$$(\eta_*(\zeta(t))[\beta\nabla\varphi(t) - \mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}(t))], \nabla\psi)_H = (q(t), \psi)_W \quad \forall \psi \in W. \tag{22}$$

The following theorem which states the existence of a unique solution to Problem VP can be proved proceeding as in [6].

Theorem 1 *Assume that* (13) − (19) *hold. Then, there exists a unique solution to Problem* VP *such that,*

$$\begin{aligned}
&\boldsymbol{u} \in C([0,T]; V), \quad \varphi \in C([0,T]; W_D),\\
&\zeta(\boldsymbol{x}, t) \in [\zeta_*, 1] \quad \textit{a.e.} \quad \boldsymbol{x} \in \Omega,\ t \in [0,T],\\
&\zeta \in H^1(0,T;Y) \cap L^2(0,T;H^2(\Omega)) \cap L^\infty(0,T;H^1(\Omega)) \cap C([0,T];H^r(\Omega))
\end{aligned}$$

for some $0 < r < 1$*.*

The proof of this theorem is based on the theory of maximal monotone operators, the Schauder fixed-point theorem and a comparison result stated in [16].

2 Numerical analysis of a fully discrete scheme

In this section a finite element algorithm is introduced for the numerical resolution of Problem VP and error estimates are obtained on the approximate solutions. In order to simplify the writing we assume, without loss of generality, that $\varphi_D = 0$ and then $W_D = W$.

The discretization of Problem VP will be done in two steps. First, we consider three finite dimensional spaces $V^h \subset V$, $W^h \subset W$ and $E^h \subset H^1(\Omega)$ which approximate the spaces V, W and $H^1(\Omega)$, respectively. Here, $h > 0$ denotes the spatial discretization parameter.

Secondly, the time derivatives are discretized by using a uniform partition of the time interval $[0,T]$, denoted by $0 = t_0 < t_1 < \ldots < t_N = T$ and let $k = T/N$ be the time step size. For a continuous function $f(t)$, let $f_n = f(t_n)$. Moreover, c denotes a positive constant which depends on the problem data but it is independent of the discretization parameters k and h.

The fully discrete approximation of Problem VP, based on a hybrid combination of the forward and the backward Euler schemes, is the following.

Problem VP^{hk}. *Find a discrete displacement field $\boldsymbol{u}^{hk} = \{\boldsymbol{u}_n^{hk}\}_{n=0}^N \subset V^h$, a discrete electric potential field $\varphi^{hk} = \{\varphi_n^{hk}\}_{n=0}^N \subset W^h$ and a discrete damage field $\zeta^{hk} = \{\zeta_n^{hk}\}_{n=0}^N \subset E^h$ such that $\zeta_0^{hk} = \zeta_0^h$ and for $n = 1, \ldots, N$,*

$$(\delta\zeta_n^{hk}, \xi^h)_Y + a(\zeta_n^{hk}, \xi^h) = (\phi(\boldsymbol{\varepsilon}(\boldsymbol{u}_{n-1}^{hk}), \zeta_{n-1}^{hk}), \xi^h)_Y \quad \forall \xi^h \in E^h, \tag{23}$$

$$(\eta_*(\zeta_n^{hk})[\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}) + \mathcal{E}^*\nabla\varphi_n^{hk}], \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\boldsymbol{u}_n^{hk}, \boldsymbol{w}^h) = (\boldsymbol{f}_n, \boldsymbol{w}^h)_V \quad \forall \boldsymbol{w}^h \in V^h, \tag{24}$$

$$(\eta_*(\zeta_n^{hk})[\beta\nabla\varphi_n^{hk} - \mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk})], \nabla\psi^h)_H = (q_n, \psi^h)_W \quad \forall \psi^h \in W^h, \tag{25}$$

where ζ_0^h is an appropriate approximation of the initial condition ζ_0, and $\boldsymbol{u}_0^{hk} \in V^h$ and $\varphi_0^{hk} \in W^h$ are the solutions to the following problems:

$$(\eta_*(\zeta_0^h)[\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_0^{hk}) + \mathcal{E}^*\nabla\varphi_0^{hk}], \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q + j(\boldsymbol{u}_0^{hk}, \boldsymbol{w}^h) = (\boldsymbol{f}_0, \boldsymbol{w}^h)_V \quad \forall \boldsymbol{w}^h \in V^h, \tag{26}$$

$$(\eta_*(\zeta_0^h)[\beta\nabla\varphi_0^{hk} - \mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_0^{hk})], \nabla\psi^h)_H = (q_0, \psi^h)_W \quad \forall \psi^h \in W^h. \tag{27}$$

Using standard arguments for nonlinear variational equations (see [11]), we deduce the existence and uniqueness of the solution to Problem VP^{hk}.

In this section, our interest is focused on the estimate of the numerical errors defined by $\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V$, $\|\varphi_n - \varphi_n^{hk}\|_W$ and $\|\zeta_n - \zeta_n^{hk}\|_Y$. We have the following.

Theorem 2 *Assume that* (13) – (19) *hold. Let $\{\boldsymbol{u}, \varphi, \zeta\}$ and $\{\boldsymbol{u}^{hk}, \varphi^{hk}, \zeta^{hk}\}$ denote the solutions to problems VP and VP^{hk}, respectively. Let us assume the following regularity conditions on the continuous solution,*

$$\begin{aligned}&\boldsymbol{u} \in C([0,T]; [W^{1,\infty}(\overline{\Omega})]^d), \quad \varphi \in C([0,T]; W^{1,\infty}(\overline{\Omega})),\\ &\zeta \in C^1([0,T]; Y) \cap C([0,T]; H^1(\Omega)).\end{aligned}$$

Therefore, we have the following error estimates for all $\boldsymbol{w}^h = \{\boldsymbol{w}_n^h\}_{n=1}^N \subset V^h$, $\psi^h = \{\psi_n^h\}_{n=1}^N \subset W^h$ and $\xi^h = \{\xi_n^h\}_{n=1}^N \subset E^h$,

$$\begin{aligned}&\max_{0\le n\le N} \Big\{ \|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V^2 + \|\varphi_n - \varphi_n^{hk}\|_W^2 + \|\zeta_n - \zeta_n^{hk}\|_Y^2 \Big\} + k\sum_{n=1}^N \|\zeta_n - \zeta_n^{hk}\|_{H^1(\Omega)}^2\\ &\quad \le c\Big(k\sum_{j=1}^N \Big[\|\zeta_j' - \delta\zeta_j\|_Y^2 + \|\boldsymbol{u}_j - \boldsymbol{u}_{j-1}\|_V^2 + \|\zeta_j - \xi_j^h\|_{H^1(\Omega)}^2 \Big] + \|\zeta_0 - \zeta_0^h\|_Y^2\\ &\qquad + k^2 + \frac{1}{k}\sum_{j=1}^{N-1} \|\zeta_j - \xi_j^h - (\zeta_{j+1} - \xi_{j+1}^h)\|_Y^2 + \|\varphi_0 - \varphi_0^{hk}\|_W^2 + \|\boldsymbol{u}_0 - \boldsymbol{u}_0^{hk}\|_V^2\\ &\qquad + \max_{1\le n\le N} \Big\{ \|\boldsymbol{u}_n - \boldsymbol{w}_n^h\|_V^2 + \|\varphi_n - \psi_n^h\|_W^2 + \|\zeta_n - \xi_n^h\|_Y^2 \Big\} \Big),\end{aligned} \tag{28}$$

where $\boldsymbol{u}_0^{hk}$ and φ_0^{hk} are the unique solutions to discrete problems (26) and (27), respectively.

PROOF.

First, we notice that the error estimate for the damage field was already done in [3]. The following estimates were obtained there for all $\xi^h = \{\xi_j^h\}_{j=1}^h \subset E^h$,

$$\begin{aligned}
\|\zeta_n - \zeta_n^{hk}\|_Y^2 + k\sum_{j=1}^n \|\nabla(\zeta_j - \zeta_j^{hk})\|_H^2 \le c\Big(k\sum_{j=1}^n \Big[&\|\boldsymbol{u}_{j-1} - \boldsymbol{u}_{j-1}^{hk}\|_V^2 \\
&+\|\zeta_{j-1} - \zeta_{j-1}^{hk}\|_Y^2 + \|\zeta_j' - \delta\zeta_j\|_Y^2 + \|\boldsymbol{u}_j - \boldsymbol{u}_{j-1}\|_V^2 + \|\zeta_j - \xi_j^h\|_{H^1(\Omega)}^2\Big] \\
&+k^2 + \max_{1\le n\le N}\|\zeta_n - \xi_n^h\|_Y^2 + \|\zeta_0 - \zeta_0^h\|_Y^2 + k\sum_{j=1}^{n-1}\|\zeta_j - \zeta_j^{hk}\|_Y^2 \\
&+\frac{1}{k}\sum_{j=1}^{n-1}\|\zeta_j - \xi_j^h - (\zeta_{j+1} - \xi_{j+1}^h)\|_Y^2\Big). \qquad (29)
\end{aligned}$$

Secondly, we turn to estimate the numerical errors on the electric potential and the displacement field. In [7] we proved the following equality,

$$\begin{aligned}
&-(\eta_*(\zeta_n^{hk})\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}), \nabla(\varphi_n - \varphi_n^{hk}))_H \\
&\quad = ((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n), \nabla(\varphi_n - \varphi_n^{hk}))_H \\
&\qquad -((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\beta\nabla\varphi_n, \nabla(\varphi_n - \varphi_n^{hk}))_H \\
&\qquad -(\eta_*(\zeta_n^{hk})\beta\nabla(\varphi_n - \varphi_n^{hk}), \nabla(\varphi_n - \varphi_n^{hk}))_H \\
&\qquad +(\eta_*(\zeta_n^{hk})\beta\nabla(\varphi_n - \varphi_n^{hk}), \nabla(\varphi_n - \psi^h))_H \\
&\qquad +((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\beta\nabla\varphi_n, \nabla(\varphi_n - \psi^h))_H \\
&\qquad -(\eta_*(\zeta_n^{hk})\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}), \nabla(\varphi_n - \psi^h))_H \\
&\qquad -((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n), \nabla(\varphi_n - \psi^h))_H \quad \forall\psi^h \in W^h. \qquad (30)
\end{aligned}$$

Substracting (20) at time $t = t_n$ and (24) for all $\boldsymbol{w}^h \in V^h$ we find that

$$\begin{aligned}
&(\eta_*(\zeta_n)[\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n) + \mathcal{E}^*\nabla\varphi_n] - \eta_*(\zeta_n^{hk})[\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}) + \mathcal{E}^*\nabla\varphi_n^{hk}], \boldsymbol{\varepsilon}(\boldsymbol{w}^h))_Q \\
&\qquad +j(\boldsymbol{u}_n, \boldsymbol{w}^h) - j(\boldsymbol{u}_n^{hk}, \boldsymbol{w}^h) = 0,
\end{aligned}$$

which leads to the following,

$$\begin{aligned}
&(\eta_*(\zeta_n)\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n) - \eta_*(\zeta_n^{hk})\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}))_Q + j(\boldsymbol{u}_n, \boldsymbol{u}_n - \boldsymbol{u}_n^{hk}) \\
&\qquad +(\eta_*(\zeta_n)\mathcal{E}^*\nabla\varphi_n - \eta_*(\zeta_n^{hk})\mathcal{E}^*\nabla\varphi_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}))_Q - j(\boldsymbol{u}_n^{hk}, \boldsymbol{u}_n - \boldsymbol{u}_n^{hk}) \\
&\quad = (\eta_*(\zeta_n)\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n) - \eta_*(\zeta_n^{hk})\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q + j(\boldsymbol{u}_n, \boldsymbol{u}_n - \boldsymbol{w}^h) \\
&\qquad +(\eta_*(\zeta_n)\mathcal{E}^*\nabla\varphi_n - \eta_*(\zeta_n^{hk})\mathcal{E}^*\nabla\varphi_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q - j(\boldsymbol{u}_n^{hk}, \boldsymbol{u}_n - \boldsymbol{w}^h)
\end{aligned}$$

for all $\boldsymbol{w}^h \in V^h$. Keeping in mind (30) and

$$\begin{aligned}
&(\eta_*(\zeta_n)\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n) - \eta_*(\zeta_n^{hk})\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q \\
&\qquad +(\eta_*(\zeta_n)\mathcal{E}^*\nabla\varphi_n - \eta_*(\zeta_n^{hk})\mathcal{E}^*\nabla\varphi_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q \\
&\quad = (\eta_*(\zeta_n^{hk})\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q \\
&\qquad +((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\mathcal{A}\boldsymbol{\varepsilon}(\boldsymbol{u}_n), \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q \\
&\qquad +(\eta_*(\zeta_n^{hk})\mathcal{E}^*\nabla(\varphi_n - \varphi_n^{hk}), \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q \\
&\qquad +((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\mathcal{E}^*\nabla\varphi_n, \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{w}^h))_Q \quad \forall \boldsymbol{w}^h \in V^h, \\
&j(\boldsymbol{u}_n, \boldsymbol{u}_n - \boldsymbol{u}_n^{hk}) - j(\boldsymbol{u}_n^{hk}, \boldsymbol{u}_n - \boldsymbol{u}_n^{hk}) \geq 0, \\
&j(\boldsymbol{u}_n, \boldsymbol{u}_n - \boldsymbol{w}^h) - j(\boldsymbol{u}_n^{hk}, \boldsymbol{u}_n - \boldsymbol{w}^h) \leq c\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V \|\boldsymbol{u}_n - \boldsymbol{w}^h\|_V, \\
&(\eta_*(\zeta_n)\mathcal{E}^*\nabla\varphi_n - \eta_*(\zeta_n^{hk})\mathcal{E}^*\nabla\varphi_n^{hk}, \boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}))_Q \\
&\quad = (\eta_*(\zeta_n^{hk})\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}), \nabla(\varphi_n - \varphi_n^{hk}))_H \\
&\qquad +((\eta_*(\zeta_n) - \eta_*(\zeta_n^{hk}))\mathcal{E}\boldsymbol{\varepsilon}(\boldsymbol{u}_n), \nabla(\varphi_n - \varphi_n^{hk}))_H,
\end{aligned}$$

using properties (13), (14), (15) and (16), taking into account the solution regularities $\boldsymbol{u} \in C([0,T];[W^{1,\infty}(\overline{\Omega})]^d)$ and $\varphi \in C([0,T];W^{1,\infty}(\overline{\Omega}))$ (which imply that $\boldsymbol{\varepsilon}(\boldsymbol{u}_n) \in [L^\infty(\overline{\Omega})]^{d\times d}$ and $\nabla\varphi_n \in [L^\infty(\overline{\Omega})]^d$), and applying several times the inequality

$$ab \leq \epsilon a^2 + (1/4\epsilon)b^2, \quad a, b, \epsilon \in \mathbb{R},$$

for some $\epsilon > 0$ small enough, after some tedious algebra we obtain the following estimates for both the electric potential and the displacement field, for all $\psi_n^h \in W^h$ and $\boldsymbol{w}_n^h \in V^h$,

$$\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V^2 + \|\varphi_n - \varphi_n^{hk}\|_W^2 \leq c\left(\|\varphi_n - \psi_n^h\|_W^2 + \|\boldsymbol{u}_n - \boldsymbol{w}_n^h\|_V^2 + \|\zeta_n - \zeta_n^{hk}\|_Y^2\right).$$

Finally, combining the previous estimates and (29), and using a discrete version of Gronwall's inequality (see [13] for details), we deduce error estimates (28). □

These error estimates are the basis for the analysis of the convergence rate of the algorithm.

As an example, let Ω be a polyhedral domain and denote by $\mathcal{T}^h$ a triangulation of $\overline{\Omega}$ compatible with the partition of the boundary $\Gamma = \partial\Omega$ into Γ_D, Γ_N and Γ_C.

Under some additional regularity conditions, the linear convergence of the algorithm can be derived, with respect to the discretization parameters h and k, which we state in the following.

Corollary 3 *Let the assumptions of Theorem 2 hold and denote by $\{\boldsymbol{u}, \varphi, \zeta\}$ and $\{\boldsymbol{u}^{hk}, \varphi^{hk}, \zeta^{hk}\}$ the respective solutions to problems VP and VP^{hk}. Let the finite element spaces V^h, W^h and E^h be composed of continuous and piecewise affine functions; that is,*

$$\begin{aligned}
&V^h = \{\boldsymbol{w}^h \in [C(\overline{\Omega})]^d \; ; \; \boldsymbol{w}^h_{|_{Tr}} \in [P_1(Tr)]^d \;\; \forall Tr \in \mathcal{T}^h, \quad \boldsymbol{w}^h = \boldsymbol{0} \quad on \quad \Gamma_D\}, \\
&W^h = \{\psi^h \in C(\overline{\Omega}) \; ; \psi^h_{|_{Tr}} \in P_1(Tr) \;\; \forall Tr \in \mathcal{T}^h, \quad \psi^h = 0 \quad on \quad \Gamma_D\}, \\
&E^h = \{\xi^h \in C(\overline{\Omega}) \; ; \; \xi^h_{|_{Tr}} \in P_1(Tr), \quad \forall Tr \in \mathcal{T}^h\},
\end{aligned}$$

where $P_1(Tr)$ represents the space of polynomial functions of global degree less or equal to 1 in Tr. Moreover, we also assume that the discrete initial condition ζ_0^h is obtained by $\zeta_0^h = \pi^h\zeta_0$, where $\pi^h : C(\overline{\Omega}) \to E^h$ is the standard finite element interpolation operator (see, e.g., [4]).

Under the additional regularity conditions

$$\begin{aligned}
&\boldsymbol{u} \in C^1([0,T];V) \cap C([0,T];[H^2(\Omega)]^d), \quad \varphi \in C([0,T];H^2(\Omega)),\\
&\zeta \in H^2(0,T;Y) \cap H^1(0,T;H^1(\Omega)) \cap C([0,T];H^1(\Omega)),
\end{aligned}$$

the numerical algorithm introduced in Problem VP^{hk} is linearly convergent; that is, there exists $c > 0$, independent of h and k, such that,

$$\max_{0\le n\le N}\{\|\boldsymbol{u}_n - \boldsymbol{u}_n^{hk}\|_V + \|\varphi_n - \varphi_n^{hk}\|_W + \|\zeta_n - \zeta_n^{hk}\|_Y\} \le c(h+k).$$

Corollary 3 is obtained using estimates (28) and taking into account the well-known approximation results and the definition of the operator π^h (see [4]),

$$\begin{aligned}
&\inf_{\xi_n^h\in E^h} \|\zeta_n - \xi_n^h\|_{H^1(\Omega)} \le ch\|\zeta\|_{C([0,T];H^2(\Omega))},\\
&\inf_{\psi_n^h\in W^h} \|\varphi_n - \psi_n^h\|_{H^1(\Omega)} \le ch\|\varphi\|_{C([0,T];H^2(\Omega))},\\
&\inf_{w_n^h\in V^h} \|\boldsymbol{u}_n - \boldsymbol{w}_n^h\|_V \le ch\|\boldsymbol{u}\|_{C([0,T];[H^2(\Omega)]^d)},\\
&\|\zeta_0 - \pi^h\zeta_0\|_Y \le ch^2\|\zeta\|_{C([0,T];H^2(\Omega))},
\end{aligned}$$

the estimate (see [12]),

$$\frac{1}{k}\sum_{j=1}^{N-1} \|\zeta_j - \xi_j^h - (\zeta_{j+1} - \xi_{j+1}^h)\|_Y \le ch\|\zeta\|_{H^1(0,T;H^1(\Omega))},$$

and the straightforward estimates (see [3]),

$$\begin{aligned}
&k\sum_{j=1}^{N} \Big[\|\zeta_j' - \delta\zeta_j\|_Y + \|\boldsymbol{u}_j - \boldsymbol{u}_{j-1}\|_V\Big] \le ck\Big(\|\zeta\|_{H^2(0,T;Y)} + \|\boldsymbol{u}\|_{C^1([0,T];V)}\Big),\\
&\|\varphi_0 - \varphi_0^{hk}\|_W + \|\boldsymbol{u}_0 - \boldsymbol{u}_0^{hk}\|_V \le ch\Big(\|\varphi\|_{C([0,T];H^2(\Omega))} + \|\boldsymbol{u}\|_{C([0,T];[H^2(\Omega)]^d)}\Big).
\end{aligned}$$

3 Numerical results

3.1 The numerical algorithm

First, the “discrete initial conditions” for the displacements and the electric potential, $\boldsymbol{u}_0^{hk}$ and φ_0^{hk}, are obtained by solving (26) and (27), simultaneously. We notice that a penalty-duality algorithm is applied for solving the nonlinearity due to the normal compliance function (see [8]).

Secondly, let the solution $(\boldsymbol{u}_{n-1}^{hk}, \varphi_{n-1}^{hk}, \zeta_{n-1}^{hk})$ at time t_{n-1} be known. We then obtain the discrete damage field at time t_n, ζ_n^{hk}, from the discrete linear variational equation

(23), which leads to a symmetric linear system and Cholesky's method is applied for its resolution.

Finally, the discrete displacement field and the discrete electric potential are obtained from the coupled equations (24) and (25), respectively. Again, the above penalty-duality algorithm is employed for solving this nonlinear problem.

The numerical scheme was implemented on a 3.2 Ghz PC using MATLAB, and a typical 2D run took about 10 minutes of CPU time.

3.2 A two-dimensional example

As a two-dimensional example, we consider an elasto-piezoelectric body, occupying the domain $\Omega = (0,6) \times (0,1.2)$, which is clamped on its left vertical boundary $\Gamma_D = \{0\} \times [0,1.2]$. We assume that no mechanical volume forces act in the body and that no volume electric charges are applied there. Moreover, we assume that a traction force, linearly increasing with respect to the x_1-variable, is acting on the horizontal upper boundary and that no surface electric charges are applied on the boundary. Finally, the body is in contact with a deformable obstacle on the horizontal lower boundary $\Gamma_C = [0,6] \times \{0\}$.

The body is assumed PZT-5A, a piezoceramic material of 6mm symmetry class with coefficients and notations detailed in [1].

The damage source function used in the numerical simulations presented below has the following form,

$$\phi(\boldsymbol{\varepsilon}(\boldsymbol{u}),\zeta) = -(\zeta-\zeta_*)_+ \quad \lambda_D \quad \frac{1-\eta_*(\zeta)}{\eta_*(\zeta)} \quad + \frac{1}{2}\lambda_U R_{q^*}(\boldsymbol{\varepsilon}(\boldsymbol{u})\cdot\boldsymbol{\varepsilon}(\boldsymbol{u}))\ {}_+ \ ,$$

where $r_+ = \max\{r,0\}$ denotes the positive part of r, λ_D and λ_U are process parameters and $R_{q^*} : \mathbb{R} \to [-q^*, q^*]$ is a truncation function given by

$$R_{q^*}(r) = \begin{cases} r & \text{if} \quad |r| \le q^*, \\ q^* & \text{otherwise.} \end{cases}$$

The normal compliance function p has the expression

$$p(u_\nu - s) = c_\nu (u_\nu - s)_+,$$

where c_ν is a deformability coefficient which represents the obstacle stiffness.

The following data were employed in these simulations:

$$\begin{aligned}
&T = 1\,s, \quad \kappa = 10^{-2}, \quad \zeta_0 = 1, \quad \lambda_D = 0.01, \quad \lambda_U = 10^3, \quad c_\nu = 10^4,\\
&\zeta_* = 0.01, \quad s = 0\,m, \quad \varphi_D = 0\,V, \quad \boldsymbol{f}_N(x_1,x_2,t) = (0, -5\times 10^4 x_1)\,N/m^2,\\
&\boldsymbol{f}_B = \boldsymbol{0}\,N/m^3, \quad q_B = 0\,C/m^3, \quad q_N = 0\,C/m^2.
\end{aligned}$$

Taking $k = 0.01$ as the time discretization parameter, the deformed mesh (amplified by 100) at final time and the initial configuration are shown in Figure 2. As expected, the contact with the obstacle is produced and the penetration into the obstacle is

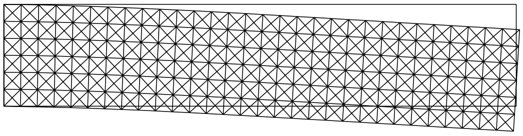


Figure 2: Deformed mesh (x100) at final time and initial configuration.

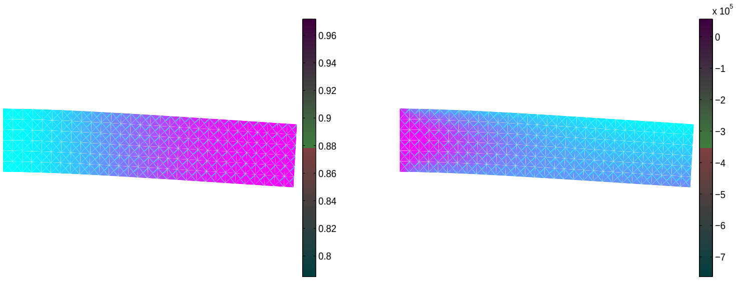


Figure 3: Damage field and electric potential at final time on the deformed mesh.

observed. In Figure 3 the damage field (left-hand side) and the electric potential (right-hand side) are plotted on the deformed mesh. The sensor effect is clearly observed: an electric potential is generated due to the deformation. Because of the vertical movement and the clamping conditions, the more damaged areas concentrate on the left part of the body.

In order to demonstrate the so-called actuator effect, a similar setting than in the above example has been considered. Now, no mechanical forces are acting and we assume a large difference of electric potential φ_D given by $\varphi_D(x_1, x_2) = 0$ if $x_1 = 0$ and $\varphi_D(x_1, x_2) = 2 \times 10^8$ if $x_1 = 6$. This example requires straightforward modifications in the writing of Problem P.

Taking again $k = 0.01$ as the time discretization parameter, the deformed mesh (amplified by 20) at final time and the initial configuration are shown in Figure 4. As expected, the contact with the obstacle is produced and the penetration into the obstacle is observed. However, since no mechanical forces were applied, we notice that this deformation is only produced by the electric displacement (i.e., the sensor effect). As a reaction to the electrical forces, the body has an bending movement. In Figure 5 the damage field (left-hand side) and the electric potential (right-hand side) are plotted on the deformed mesh. Due to the resulting deformation, the most damaged areas are located where the body bends and also near to the clamped part. Moreover, since a large difference of electric potential is assumed, the electric potential seems to be constant through the horizontal direction.

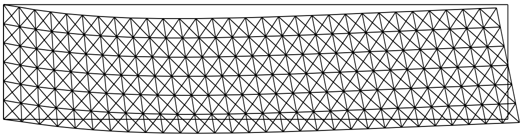


Figure 4: Deformed mesh (x20) at final time and initial configuration.

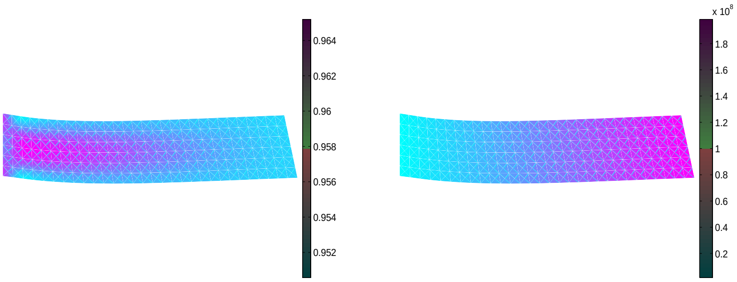


Figure 5: Damage field and electric potential at final time on the deformed mesh.

Acknowledgements

This work was partially supported by the Ministerio de Educación y Ciencia (Project MTM2006-13981).

References

[1] M. Barboteu, J.R. Fernández, Y. Ouafik, *Numerical analysis of two frictionless elasto-piezoelectric contact problems*, J. Math. Anal. Appl. **329** (2008) 905–917.

[2] R.C. Batra, J.S. Yang, *Saint-Venant's principle in linear piezoelectricity*, J. Elasticity **38** (1995) 209–218.

[3] M. Campo, J.R. Fernández, K.L. Kuttler, M. Shillor, *Quasistatic evolution of damage in an elastic body: numerical analysis and computational experiments*, Appl. Numer. Math. 2007; **57**(9) (2007) 975–988.

[4] P.G. Ciarlet, *The finite element method for elliptic problems*, In: P.G. Ciarlet and J.L. Lions, (Eds.), *Handbook of Numerical Analysis*, Vol. II, North Holland, 1991, pp. 17–352.

[5] G. Duvaut, J.L. Lions, *Inequalities in mechanics and physics*, Springer-Verlag, Berlin, 1976.

[6] J.R. Fernández, K.L. Kuttler, *An existence and uniqueness result for an elasto-piezoelectric problem with damage*, Math. Models Methods Appl. Sci. (in press).

[7] J.R. Fernández, R. Martínez, G.E. Stavroulakis, *Numerical analysis of an elasto-piezoelectric problem with damage*, Internat. J. Numer. Meth. Engrg. (to appear).

[8] J.R. Fernández, M. Sofonea, J.M. Viaño, *A frictionless contact problem for elastic-viscoplastic materials with normal compliance: numerical analysis and computational experiments*, Numer. Math. **90** (2002) 689–719.

[9] M. Frémond, B. Nedjar, *Damage in concrete: the unilateral phenomenon*, Nuclear Engineering Design **156** (1995) 323–335.

[10] M. Frémond, B. Nedjar, *Damage, gradient of damage and principle of virtual work*, International Journal of Solids and Structures **33**(8) (1996) 1083–1103.

[11] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer-Verlag, New York, 1984.

[12] W. Han, M. Shillor, M. Sofonea, *Variational and numerical analysis of a quasistatic viscoelastic problem with normal compliance, friction and damage,* Journal Comput. Appl. Math. **137** (2001) 377–398.

[13] W. Han, M. Sofonea, *Quasistatic contact problems in viscoelasticity and viscoplasticity*, American Mathematical Society-Intl. Press, 2002.

[14] T. Ideka, *Fundamentals of piezoelectricity*, Oxford University Press, Oxford, 1990.

[15] A. Klarbring, A. Mikelić, M. Shillor, *Frictional contact problems with normal compliance*, Internat. J. Engrg. Sci. **26** (1988) 811–832.

[16] K.L. Kuttler, *Quasistatic evolution of damage in an elastic-viscoplastic material*, Electron. J. Differential Equations **147** (2005) 1–25.

[17] J. Lemaitre, R. Desmorat, *Engineering damage mechanics: ductile, creep, fatigue and brittle failures.* Springer-Verlag, 2005.

[18] R. Liebe, P. Steinmann, A. Benallal, *Theoretical and numerical aspects of a thermodynamically consistent framework for geometrically linear gradient damage*, Comput. Methods Appl. Mech. Engng. **190** (2001) 6555–6576.

[19] J.A.C. Martins, J.T. Oden, *Existence and uniqueness results for dynamic contact problems with nonlinear normal and friction interface laws*, Nonlinear Anal. **11** (1987) 407–428.

[20] C. Miehe, *Discontinuous and continuous damage evolution in Ogden-type large-strain elastic materials*, Eur. J. Mech. A Solids **14**(3) (1995) 697–720.

[21] R.D. Mindlin, *Continuum and lattice theories of influence of electromechanical coupling on capacitance of thin dielectric films*, Internat. J. Solids Structures **4** (1969) 1197–1213.

[22] R. Turbé, G.A. Maugin, On the linear piezoelectricity of composite materials, *Math. Methods Appl. Sci.* **14**(6) (1991) 403–412.

[23] Y. Xinhua, C. Chuanyao, H. Yuantai, W. Cheng, *Combined damage fracture criteria for piezoelectric ceramics*, Acta Mechanica Solida Sinica **18**(1) (2005) 21–27.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

A dynamical model of parallel computation on bi-infinite time-scale

Wit Foryś[1], Juan Louis García Guirao[2] and Piotr Oprocha[3]

[1] *Institute of Computer Science, Jagiellonian University, Nawojki 11, 30-072 Kraków, Poland*

[2] *Departamento de Matemática Aplicada y Estadística, Hospital de Marina, 30203-Cartagena (Región de Murcia), Spain. , Universidad Politécnica de Cartagena*

[3] *Faculty of Applied Mathematics, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland*

emails: `forysw@ii.uj.edu.pl` , `Juan.Garcia@upct.es`, `oprocha@agh.edu.pl`

Abstract

The aim of this article is to construct a dynamical model of parallel computation on bi-infinite time-scale. Our approach is similar to two-sided symbolic dynamics, however bi-infinite sequences are transformed to bi-infinite graphs for a suitable description of parallelism and concurrency.

1 Introduction

A computational process executed by some system can be considered, from the general point of view, as a motion and thus can be described as a dynamical system. Then it can be a subject of analysis using all tools of dynamical system theory. The process of computation on one processor can be visualized as a sequence of successive executions of commands defined over a finite set of instructions. It is exactly the situation in symbolic dynamics, when a sequence of symbols from a finite alphabet is shifted from the right to the left. If our computation schedule is large (i.e. we have a long list of instructions to be executed) then we can agree that our computation schedule is represented by a bi-infinite sequence of symbols (i.e. sequence indexed by $\mathbb{Z}$). In such a case the index 0 points out the instruction currently executed. Notice that this kind of generalization (i.e. description of a problem by infinite models instead of finite ones) is very natural and common (e.g. most of models in engineering assume that there are infinitely many atoms in given solid etc.). The same way we assume that the tape in Turing model is infinite, because we do not want to restrict the length of computations.

Modelling a dynamical behavior by a sequence of symbols is not a new idea. It was originally applied by Hadamard in the late XIX century.

The problem of interrelations between dynamics of computation and computation process itself, perceived in theory of computing firstly, is a subject of a research which is considered from several points of view [1, 6, 9, 10, 11, 12]. Nevertheless we are still far from the full understanding of it. This is caused, mostly by the fact, that many questions about dynamics in proposed till now models are undecidable.

In contrast to the above mentioned development of sequential models, not much is known about dynamical aspects of parallel computation. In our research on relations between the dynamics of sequential computations and their parallel counterparts we introduced in [5] a framework, joining the main ideas of Symbolic Dynamics and theory of Traces, which is used to model parallel computations. However disadvantage of this model is that in each step we loose some information about the path of computation which is responsible for the current state of the system.

In the paper we will construct a framework which is similar to bi-infinite sequences in symbolic dynamics. We have to develop a technique which will transform a linear computation schedule (represented by bi-infinite sequence) into a new object which represents schedule of parallel computing. In particular, the symbol at index 0 and the future of computation (represented by symbols at indices from $\mathbb{N}$) must be related to the future in a schedule in our model of parallel computing. In a similar way as in [5] we undertake this research problem basing on theory of traces, as a tool of description of parallelism and using methods of symbolic dynamics. However we will have to develop some additional objects which will be suitable for bi-infinite time scale.

The notion of a trace was introduced by P. Cartier and D. Foata in [2] in a combinatorial context. Later, with a great success, this notion was applied by A. Mazurkiewicz [8] to model the concepts of parallelism and concurrency. Presently, there are numerous research papers, monographs and textbooks on trace theory (see e.g. [3, 4]) however they focus, mostly, on combinatorial aspects of modelled processes. Dynamical properties of parallel computational processes described by traces and sequential computations in relation to their parallel counterparts in a space of traces are still challenging problems.

In our research we will focus on dynamical model of parallel computation on bi-infinite time-scale. We have to extend our framework of [5] to model computations with the specified past. In fact we introduce bi-infinite traces to model such computations and extend a notion of a trace shift map to cover this bi-infinite case. As we will see, the situation is even more complex than this of [5] and introduced shift operator Φ (on bi-infinite traces) give rise to much richer dynamics. The further study of the properties of Φ seems to be a challenging problem. We hope that our framework will generate a deeper insight into the dynamics of parallel computation and will be motivating for a further development in the theory of computing.

2 Definitions and notations

Let Σ^* denote the set of all finite words over an alphabet Σ. Σ^*. We denote by $I \subset \Sigma \times \Sigma$ a symmetric and irreflexive relation called the independence relation. I induces a congruence $\sim_I$ on Σ^*. Considered a word as a sequence of actions denoted by letters and executed on some sets of resources we can interpret the independency of a and b as a possibility of a parallel execution of these two actions. The complement of I, that is $(\Sigma \times \Sigma) \setminus I$ is called the dependence relation and is denoted by D. The quotient $\Sigma^* / \sim_I$ is the free partially commutative monoid and is denoted by $\mathbb{M}(\Sigma, I)$ or by $\mathbb{M}(\Sigma, D)$. The elements of $\mathbb{M}(\Sigma, I)$ are called traces. A word $w \in \Sigma^*$ is in Foata normal form, if it is the empty word or if there exist an integer $n > 0$ and nonempty words $v_1, ..., v_n \in \Sigma^+$ such that the following three conditions are satisfied: $w = v_1 .....v_n$,; for any $i = 1, ..., n$ the word v_i is a catenation of pairwise independent letters and v_i is minimal with respect to the lexicographic ordering; for any $i = 1, ..., n-1$ and for any letter $a \in alph(v_{i+1})$ there exists a letter $b \in alph(v_i)$ such that $(a, b) \in D$.

A dependence graph $G = [V, E, \lambda]$ over (Σ, D) consists of

1. V, a countable set of vertices
2. $E \subset V \times V$, an edge relation such that the directed graph (V, E) is acyclic and the induced partial ordering is well founded
3. $\lambda : V \to \Sigma$, a vertex labelling function such that $(\lambda(x), \lambda(y)) \in D$ if and only if $(x, y) \in E \cup E^{-1} \cup \Delta_V$ where $\Delta_V = \{(x, x) \ : \ x \in V\}$.

The set of all dependence graphs is a monoid denoted by $\mathbb{G}(\Sigma, D)$ and $\mathbb{M}(\Sigma, I)$ is isomorphic to a submonoid of $\mathbb{G}(\Sigma, D)$ consisted of all finite graphs. We have a natural morphism $\phi_{\mathbb{G}} : \Sigma^* \longrightarrow \mathbb{G}(\Sigma, D)$. We may extend $\phi_{\mathbb{G}}$ from finite to infinite words, putting for $w = w_1, w_2, ...$ the dependence graph $\phi_{\mathbb{G}}(w) = [V, E, \lambda]$ where $V = \mathbb{N}$ and $\lambda(i) = w_i$ for any $i \in \mathbb{N}$. There exists an arrow $(i, j) \in E$, if and only if $i < j$ and $(w_i, w_j) \in D$. The image $\phi_{\mathbb{G}}(w) \in \mathbb{G}$ for any $w \in \Sigma^\infty$ is called a real trace. The family of all real traces is denoted by $\mathbb{R}(\Sigma, D)$ or $\mathbb{R}(\Sigma, I)$. If $t = sq$ then s is called a prefix of t and denoted by $s \leqslant t$. Now we introduce a topology in $\mathbb{R}(\Sigma, I)$ defining a metric as follows.

$$d_{\mathbb{R}}(s, t) = \begin{cases} 2^{-l_{\mathbb{R}}(s,t)} & if \;\; x \neq y \\ 0 & otherwise \end{cases}$$

where
$l_{\mathbb{R}}(s, t) = \sup\{n \in \mathbb{N} \ : \ \forall p \in \mathbb{M}(\Sigma, I), \ | p | \leq n, \ \ p \leq s \leftrightarrow p \leq t\}$.

3 Two-sided real traces

Definition 1 *Let (Σ, D) be a dependence alphabet. A dependence graph with dot $G = [V, E, \lambda]$ consists of*

1. *V, a countable set of vertices*

2. $\odot$*, an additional vertex* $\odot \notin V$ *and* $\widetilde{V} = V \cup \{\odot\}$

3. $E \subset \widetilde{V} \times \widetilde{V}$*, an edge relation such that the directed graph* $(\widetilde{V}, E)$ *is acyclic*

4. $\lambda : V \to \Sigma$*, a vertex labelling function such that* $(\lambda(x), \lambda(y)) \in D$ *if and only if* $(x, y) \in E \cup E^{-1} \cup \Delta_V \cap V \times V$*, where* $\Delta_V = \{(x, x) \ : \ x \in V\}$

5. $V = L(G) \cup R(G)$ *and* $L(G) \cap R(G) = \emptyset$*, where*

$$\begin{aligned} L(G) &= \{v \in V \ : \ \text{there is a path from } v \text{ to } \odot \text{ in } G\} \\ R(G) &= \{v \in V \ : \ \text{there is a path from } \odot \text{ to } v \text{ in } G\} \end{aligned}$$

The set of all dependence graphs with dot is denoted by $\mathbb{G}^\bullet(\Sigma, D)$. It could be converted into a monoid with the empty graph $1 = [\{\odot\}, \emptyset, \emptyset]$ as the neutral element of a concatenation of graphs defined as follows. For dependence graphs with dot $G_1 = [V_1 \cup \{\odot\}, E_1, \lambda_1]$ and $G_2 = [V_2 \cup \{\odot\}, E_2, \lambda_2]$ we put

$$G_1 \cdot G_2 = [V_1 \cup \{\odot\}, E_1, \lambda_1] \cdot [V_2 \cup \{\odot\}, E_2, \lambda_2] = [(V_1 \dot\cup V_2 \cup \{\odot\}), E_1 \dot\cup E_2 \dot\cup A, \lambda_1 \dot\cup \lambda_2]$$

where $A = \{(x, y) \in L(G_2) \times L(G_1) \cup R(G_1) \times R(G_2) \ : \ (\lambda_1(x), \lambda_2(y)) \in D\}$

Let us define a mapping $\varphi^R_{\mathbb{G}} : \Sigma \to \mathbb{G}^\bullet(\Sigma, D)$ putting for a letter $a \in \Sigma$ the two vertices graph $\varphi^R_{\mathbb{G}}(a) = [\{a, \odot\}, \{(\odot, a)\}, \lambda]$ where $\lambda(a) = a$. This mapping can be extended to a morphism $\varphi^R_{\mathbb{G}} : \Sigma^* \to \mathbb{G}^\bullet(\Sigma, D)$ by putting $\varphi^R_{\mathbb{G}}(ua) = \varphi^R_{\mathbb{G}}(u)\varphi^R_{\mathbb{G}}(a)$ for any $a \in \Sigma$ and $u \in \Sigma^+$. Similarly we define a mapping $\varphi^L_{\mathbb{G}} : \Sigma^* \to \mathbb{G}^\bullet(\Sigma, D)$ but now $\varphi^L_{\mathbb{G}}(a) = [\{a, \odot\}, \{(a, \odot)\}, \lambda]$, $\lambda(a) = a$ and $\varphi^L_{\mathbb{G}}(ua) = \varphi^L_{\mathbb{G}}(a) \cdot \varphi^L_{\mathbb{G}}(u)$ for any $a \in \Sigma$ and $u \in \Sigma^+$. Denote by $u.v$ a pair of words and let $S(\Sigma) = \{u.v \ : \ u, v \in \Sigma^*\}$. Define a map $\varphi^\bullet_{\mathbb{G}} : S(\Sigma) \to \mathbb{G}^\bullet(\Sigma, D)$ putting $\varphi^\bullet_{\mathbb{G}}(u.v) = \varphi^L_{\mathbb{G}}(u) \cdot \varphi^R_{\mathbb{G}}(u)$.

We can extend the mapping $\varphi^\bullet_{\mathbb{G}}$ from two-sided to bi-infinite words. For $w = \ldots, w_{-1}, w_0, w_1, \ldots$ the dependence graph with dot $\varphi^\bullet_{\mathbb{G}}(w) = [V \cup \{\odot\}, E, \lambda]$ is defined as follows. We put $V = \mathbb{Z}$ and $\lambda(i) = w_i$ for any $i \in \mathbb{Z}$. There exists an arrow $(u, v) \in E$, if and only if

1. $u, v \in V$, $u < v < 0$ or $0 \leq u < v$, $(w_u, w_v) \in D$,
2. $u < 0$ and $v = \odot$,
3. $u = \odot$ and $v \geq 0$.

That way we have defined bi-infinite traces. The image $\varphi^\bullet_{\mathbb{G}}(w)$ for any $w \in S(\Sigma) \cup \Sigma^{\mathbb{Z}}$ is called a two-sided real trace. The family of all real two-sided traces is denoted by $\mathbb{R}^\bullet(\Sigma, D)$. We denote the set of finite two-sided traces (with finitely many vertices) by $\mathbb{M}^\bullet(\Sigma, I)$. The set of bi-infinite real traces is denoted $\mathbb{R}^{\omega\bullet\omega}(\Sigma, D)$.

If $t \in \mathbb{M}^\bullet(\Sigma, I)$ then define its *length* to be the integer $|t| = \max\{\#L(t), \#R(t)\}$. If there exist traces s, t, r such that $t = s \cdot r$ then we write $s \leq t$ and call s a (two-sided) prefix of t. If $s \leq t$ and $t \leq s$ then we write $s = t$, otherwise $s \neq t$.

Let $G = [V, E, \lambda] \in \mathbb{R}^\bullet(\Sigma, D)$. For any $v \in R(G)$ we denote by $|v|$ the length of the longest path from $\odot$ to v. If $v \in L(G)$ then $|v|$ is the length of the longest path from v to $\odot$. For any integer $n > 0$ we define $F_n^R(G) = w$ where the word w is catenation of letters $\{\lambda(v) \,:\, v \in R(G), |v| = n\}$ in order defined by the path in G starting in $\odot$ and $F_n^L(G)$ is the catenation of letters from $\{\lambda(v) \,:\, v \in L(G), |v| = n\}$. We also define $F_0^L(G) = F_0^R(G) = 1$. To unify the notation we define

$$F_i(t) = \begin{cases} F_i^R(t) & , i \geq 0 \\ F_i^L(t) & , i < 0. \end{cases}$$

For any two two-sided traces $s, t \in \mathbb{R}^\bullet(\Sigma, D)$, $s \neq t$ we define

$$\begin{aligned} l_{\text{pref}}(s,t) &= \max\{n \,:\, r \leq s \Leftrightarrow r \leq t \text{ for every } r \in \mathbb{M}^\bullet(\Sigma, I) \text{ with } |r| \leq n\} \\ l_{\text{fnf}}(s,t) &= \max\{n \,:\, F_i(s) = F_i(t) \text{ for } i = -n, \ldots, n\} \end{aligned}$$

and that way we introduce two metrics on $\mathbb{R}^\bullet(\Sigma, D)$ called *prefix metric* and *Foata normal form metric* respectively:

$$d_{\text{pref}}(s,t) = \begin{cases} 2^{-l_{\text{pref}}(s,t)} & , s \neq t \\ 0 & , s = t \end{cases} \quad , \quad d_{\text{fnf}}(s,t) = \begin{cases} 2^{-l_{\text{fnf}}(s,t)} & , s \neq t \\ 0 & , s = t \end{cases}$$

Proposition 2 *Metrics d_{pref} and d_{fnf} are uniformly equivalent.*

Proposition 3 *Monoids $\mathbb{M}^\bullet(\Sigma, I)$ and $\mathbb{M}(\Sigma, D) \times \mathbb{M}(\Sigma, D)$ are isomorphic.*

Proposition 4 *If we endow the space $\mathbb{R}(\Sigma, D) \times \mathbb{R}(\Sigma, D)$ with the metric $d((x_1, x_2), (y_1, y_2)) = \max\{d_{\mathbb{R}}(x_1, y_1), d_{\mathbb{R}}(x_2, y_2)\}$ then we obtain a metric space homeomorphic to $(\mathbb{R}^\bullet(\Sigma, D), d_{fnf})$. The same is true for $\mathbb{R}^{\omega\bullet\omega}(\Sigma, I)$ and $\mathbb{R}^\omega(\Sigma, I) \times \mathbb{R}^\omega(\Sigma, I)$.*

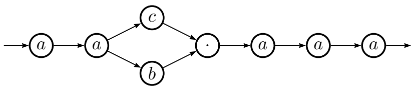
4 Shifts on bi-infinite traces

Given $t \in \Phi$ and $i \in \mathbb{Z}$, $i \neq 0$ let $\psi(i, t)$ be a bi-infinite trace obtained from t by removing all vertices with $|v| = i$. We define a map $\Phi : \mathbb{R}^{\omega\bullet\omega}(\Sigma, I) \to \mathbb{R}^{\omega\bullet\omega}(\Sigma, I)$ which is an analogue of σ on $\mathcal{A}^{\mathbb{Z}}$. Given $t \in \mathbb{R}^{\omega\bullet\omega}(\Sigma, I)$ the map Φ will shift $F_1(t)$ from the right to the left side of the vertex $\odot$. Strictly speaking

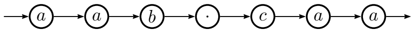
$$\Phi(t) = \varphi_L(F_1(t)) \cdot \psi(1, t).$$

The mapping Φ seems to be similar to σ, however it is only the first impression. The most important difference is that Φ is not invertible. In fact it is neither injective nor surjective.

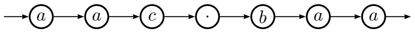
Example 5 *Let $\Sigma = \{a, b, c\}$ and let the relation I be represented by the undirected graph $b - c$. Note that in that case the bi-infinite trace:*



can be obtained as the image by Φ *of bi-infinite trace*



or



Additionally observe that none of two graphs above can be obtained as an image of Φ*.*

Proposition 6 *The map* Φ *is continuous.*

Proposition 7 *The set* $X = \bigcap_{n=0}^{\infty} \Phi^n(\mathbb{R}^{\omega\bullet\omega}(\Sigma, I))$ *is closed and* Φ *invariant (i.e.* $\Phi(X) = X$*). It is the largest set (in the sense of inclusion) with these properties. The case* $X = \mathbb{R}^{\omega\bullet\omega}(\Sigma, I)$ *holds iff* $I = \emptyset$*. In that case* $(\mathbb{R}^{\omega\bullet\omega}(\Sigma, I), \Phi)$ *is equivalent to* $(\Sigma^{\mathbb{Z}}, \sigma)$*.*

Proposition 8 *Let* t *be a bi-infinite trace. If* $F_{-1}(t)F_1(t) \sim_I F_{-1}(\Phi(t))$ *then* $t \notin X$*.*

Proposition 9 *Assume that* $I \neq \emptyset$*. Then there exists a shift* $X \subset \Sigma^{\mathbb{Z}}$ *such that* $\varphi^{\bullet}_{\mathbb{G}}(X)$ *is not closed. In particular the map* $\varphi^{\bullet}_{\mathbb{G}}$ *is not continuous.*

References

[1] O. Bournez and M. Cosnard. On the computational power of dynamical systems and hybrid systems. *Theoret. Comput. Sci.*, 168(2):417–459, 1996.

[2] P. Cartier and D. Foata. *Problèmes combinatories de commutation et réarrangements*. Lecture Notes in Mathematics. Springer-Verlag, 1969.

[3] V. Diekert and Y. Métivier. Partial commutation and traces. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 3, chapter 8, pages 457–534. Springer-Verlag, 1997.

[4] V. Diekert and G. Rozenberg, editors. *The book of traces*. World Scientific Publishing Co. Inc., River Edge, NJ, 1995.

[5] W. Foryś and P. Oprocha, *Infinite Traces and Symbolic Dynamics*, Theory of Computing Systems, to apperar

[6] P. Kurka. On topological dynamics of Turing machines. *Theoret. Comput. Sci.*, 174(1-2):203–216, 1997.

[7] M. Kwiatkowska, *A metric for traces*, Inform. Process. Lett., **35** (1990), 129–135.

[8] A. Mazurkiewicz. Concurrent program schemes and their interpretations. *DAIMI Rep. Aarhus University*, 78:1–45, 1977.

[9] C Moore. Unpredictability and undecidability in dynamical systems. *Phys. Rev. Lett.*, 64(20):2354–2357, 1990.

[10] H. T. Siegelmann. *Neural networks and analog computation.* Progress in Theoretical Computer Science. Birkhäuser Boston Inc., Boston, MA, 1999. Beyond the Turing limit.

[11] H. T. Siegelmann and S. Fishman. Analog computation with dynamical systems. *Physica D*, 120(1-2):214235, 1998.

[12] S. Wolfram. *A new kind of science.* Wolfram Media, Inc., Champaign, IL, 2002.

*Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.*

Positive topological entropy of Coupled Map Lattice

Juan Luis García Guirao[1] **and Marek Lampart**[2]

[1] *Departamento de Matemática Aplicada y Estadística, Hospital de Marina, 30203-Cartagena (Región de Murcia), Spain. , Universidad Politécnica de Cartagena*

[2] *Department of Applied Mathematics. VŠB, 17. listopadu 15/2172, 708 33 Ostrava, Czech Republic., Technical University of Ostrava*

emails: `juan.garcia@upct.es`, `marek.lampart@vsb.cz`

Abstract

In this paper we present a lattice dynamical system stated by K. Kaneko in [Phys. Rev. Lett., **65**, 1391-1394, 1990] which is related to the Belusov-Zhabotinskii reaction. We prove that this CML (Coupled Map Lattice) system has positive topological entropy for zero coupling constant.

Key words: coupled map lattice, positive topological entropy
MSC 2000: 37L60, 37B99, 37B10

1 Introduction

Classical Discrete Dynamical Systems (DDS's), i.e., a couple composed by a space X (usually compact and metric) and a continuous self-map ψ on X, have been highly considered in the literature (see e.g., [BC] or [De]) because are good examples of problems coming from the theory of Topological Dynamics and model many phenomena from biology, physics, chemistry, engineering and social sciences (see for example, [Da], [KO], [Pu] or [Po]). In most cases in the formulation of such models ψ is a C^∞, an analytical or a polynomial map.

Coming from chemical engineering applications, such a digital filtering, imaging and spatial vibrations of the elements which compose a given chemical product, a generalization of DDS's have recently appeared as an important subject for investigation, we mean the so called *Lattice Dynamical Systems* or *1d Spatiotemporal Discrete Systems*. In the next section we provide all the definitions. To show the importance of these type of systems, see for instance [ChF].

To analyze when one of this type of systems have a complicated dynamics or not by the observation of one topological dynamics property is an open problem. The aim of the present paper is to show by using the notion of *topological entropy* to characterize the dynamical complexity of coupled lattice systems showing the existence of positive topological entropy for zero coupling constant. We present some other problems for the future.

Let us recall the notion of Positive topological entropy which is known to topological chaos.

An attempt to measure the complexity of a dynamical system is based on a computation of how many points are necessary in order to approximate (in some sense) with their orbits all possible orbits of the system. A formalization of this intuition leads to the notion of topological entropy of the map f, which is due to Adler, Konheim and McAndrew [AKM]. We recall here the equivalent definition formulated by Bowen [B], and independently by Dinaburg [Di]: the *topological entropy* of a map f is a number $h(f) \in [0, \infty]$ defined by

$$h(f) = \lim_{\varepsilon \to 0} \limsup_{n \to \infty} \#E(n, f, \varepsilon),$$

where $E(n, f, \varepsilon)$ is a (n, f, ε)–span with minimal possible number of points, i.e., a set such that for any $x \in \mathbb{X}$ there is $y \in E(n, f, \varepsilon)$ satisfying $d(f^j(x), f^j(y)) < \varepsilon$ for $1 \leq j \leq n$.

A map f is *topologically chaotic* (briefly, PTE) if its topological entropy $h(f)$ is positive.

2 Notation and basic construction

The state space of LDS (Lattice Dynamical System) is the set

$$\mathcal{X} = \{x | \ x = \{x_i\}, \ x_i \in \mathbb{R}^d, \ i \in \mathbb{Z}^D, \ \| x_i \| < \infty\},$$

where $d \geq 1$ is the dimension of the range space of the map of state x_i, $D \geq 1$ is the dimension of the lattice and the l^2 norm $\| x \|_2 = (\Sigma_{i \in \mathbb{Z}^D} \mid x_i \mid^2)^{1/2}$ is usually taken ($\mid x_i \mid$ is the length of the vector x_i).

We deal with the following 1d-LD CML (Coupled Map Lattice) system which was stated by K. Kaneko in [K] (for more details see for references therein) and it is related to the Belusov-Zhabotinskii reaction (see [KO] and for experimental study of chemical turbulence by this method [HGS], [HOY], [HHM]):

$$x_n^{m+1} = (1 - \epsilon) f(x_n^m) + \epsilon/2 [f(x_{n-1}^m) - f(x_{n+1}^m)], \tag{1}$$

where m is discrete time index, n is lattice side index with system size L (i.e. $n = 1, 2, \ldots L$), ϵ is coupling constant and $f(x)$ is the *unimodal map* on the unite closed interval $I = [0, 1]$, i.e. $f(0) = f(1) = 0$ and f has unique critical point c with $0 < c < 1$ such that $f(c) = 1$. For simplicity we will deal with so called "tent map", defined by

$$f(x) = \begin{cases} 2x, & x \in [0, 1/2), \\ 2 - 2x, & x \in [1/2, 1]. \end{cases}$$

In general, one for the following periodic boundary conditions of the system (1) is assumed:

1. $x_n^m = x_{n+L}^m$,
2. $x_n^m = x_n^{m+L}$,
3. $x_n^m = x_{n+L}^{m+L}$,

standardly, the first case of the boundary conditions is used.

The equation (1) was studied by many authors, mostly experimentally or semi-analytically then analytically. The first paper with analytic results is [ChL], where it was proved that this system is chaotic in the sense of Li and Yorke.

We consider, as an example the 2-element one-way coupled logistic lattice (OCLL, see [KW]) $H : I^2 \to I^2$ written as

$$\begin{aligned} x_1^{m+1} &= (1-\epsilon) f(x_1^m) + \epsilon f(x_2^m), \\ x_2^{m+1} &= \epsilon f(x_1^m) + (1-\epsilon) f(x_2^m), \end{aligned} \tag{2}$$

where f is the tent map.

The following construction is similar to the [BCP]. Since the critical point for the tent map is equal to 1/2 we can divide the interval I into two sets $P_1 = [0, 1/3)$ and $P_2 = (2/3, 1]$ and get a family $\mathcal{P} = \{P_1, P_2\}$. Then each point $x_0 \in \Lambda_1$ can be represented as an infinite symbol sequence $C_1(x_0) = \alpha = a_1 a_2 a_3 \dots$ where Λ_1 is Cantor ternary set and

$$a_n = \begin{cases} 0 & \text{if} \quad f^n(x_0) \in P_1, \\ 1 & \text{if} \quad f^n(x_0) \in P_2. \end{cases}$$

Returning to (2) we can divide its range set into four sets $\mathcal{P} = \{P_1^1, P_2^1, P_1^2, P_2^2\}$ where the upper index corresponds to the x_1 coordinate and x_2 to the lower one. Then again each point $p \in \Lambda_2$ can be encrypted as an infinite symbol sequence $C_2(p) = \alpha = a_1 a_2 a_3 \dots$ where Λ_2 is 2-dimensional Cantor ternary set [1] and

$$a_n = \begin{cases} 0 & \text{if} \quad H^n(p) \in P_1^1, \\ 1 & \text{if} \quad H^n(p) \in P_2^1, \\ 2 & \text{if} \quad H^n(p) \in P_1^2, \\ 3 & \text{if} \quad H^n(p) \in P_2^2. \end{cases}$$

Now, we denote the *k-shift* operator σ_k on k symbol alphabet, defined by $\sigma_k : \Sigma_k \to \Sigma_k$ and $\sigma_k(a_1 a_2 a_3 \dots) = a_2 a_3 \dots$ where $\Sigma_k = \{\alpha \mid \alpha = a_1 a_2 a_3 \dots \text{and } a_i \in \{1, 2, \dots k\}\}$, so the effect of this operator is to delete the first symbol of the sequence α.

We can observe that Λ_2 is invariant[2] subset of the range space of the system (2) and that each its point is encoded by exactly one point from Σ_4, for $\epsilon = 0$. So, by [F] the shift operator σ_4 acts on Σ_4 exactly as (2) on Λ_2, for $\epsilon = 0$.

[1] by *n-dimensional Cantor set* we mean the Cantor set constructed as subset of $\mathbb{R}^n$

[2] a set M is invariant for the map f if $f(M) \subset M$

3 Main result

We say that two dynamical systems (X, f) and (Y, g) are *topologically conjugated* if there is a homeomorphism $h : X \to Y$ such that $h \circ f = g \circ h$ (the diagram commutes), such homeomorphism is called *conjugacy*. Then:

Proposition 1 ([W]) *If (X, f) and (Y, g) are topologically conjugated systems then $h(f) = h(g)$.*

For the proof of the main result we also use well known result:

Proposition 2 ([W]) *Let σ_k be the k-shift operator. Then $h(\sigma_k) = k \log 2$.*

Theorem 1 *The system*

$$x_n^{m+1} = (1-\epsilon) f(x_n^m) + \epsilon/2[f(x_{n-1}^m) - f(x_{n+1}^m)],$$

has positive topological entropy for $\epsilon = 0$. Moreover, its entropy equals to $L \log 2$.

Proof 1 *By the construction of the Section 2 it follows that the 2–dimensional system (1) contains 2-dimensional Cantor set which is conjugated (see, e.g. [F]) to the shift space Σ_4 by the conjugacy map C_2, for $\epsilon = 0$. Then by Proposition 1 the system has topological entropy equal to the entropy of σ_4. Consequently, by Proposition 2 its entropy is $2 \log 2$.*

To the end of the proof, it suffice to note, that the construction of the Section 2 can be generalized to the L-dimensional systems. Such system will be conjugated to the 2^L-shift by C_L conjugacy and by the same arguments, as in the paragraph above, its entropy equals to $L \log 2$.

4 Concluding remarks

The proof of the main result can be done in an alternative way. For zero coupling constant it is obvious that each lattice side contains a subsystem conjugated to (Σ_2, σ_2). Then the system (1) contains subsystem conjugated to the L-times product of (Σ_2, σ_2) and by $h(\underbrace{\sigma_2 \times \cdots \times \sigma_2}_{L}) = Lh(\sigma_2)$ (see, e.g. [W]) the assertion follows.

For non-zero coupling constants the dynamical behaviour of the system (1) is more complicated. The first question is how the invariant subsets of phase space looks like? Secondly, is the set of periodic points dense in the range space? The answer for this question will be nontrivial. Similar system was studied in [BGLL] and there was used the method of resultants to prove existence of periodic points of higher order. The same concept like in [BGLL] should be used.

References

[AKM] R.L. Adler, A.G. Konheim and M.H. McAndrew, *Topological entropy*, Trans. Amer. Math. Soc., **114**, 309-319, 1965.

[BC] L.S. Block and W.A. Coppel, *Dynamics in One Dimension*, Springer Monographs in Mathematics, Springer-Verlag, 1992.

[BCP] E. Bollt, N.J. Corron and S.D. Pethel, *Symbolic dynamics of coupled map lattice*, Phys. Rew. Lett., **96**, 1-4, 2006.

[B] R. Bowen, *Entropy for group endomorphisms and homogeneous spaces*, Trans. Amer. Math. Soc., **153**, 401-414, 1971.

[BGLL] F. Balibrea, J.L. Garca Guirao, M. Lampart and J. Llibre, *Dynamics of a Lotka-Volterra map*, Fund. Math. 191 **3**, 265279, 2006.

[HHM] J.L. Hudson, M. Hart and D. Marinko, *An experimental study of multiplex peak periodic and nonperiodic oscilations in the Belusov-Zhabotinskii reaction*, J. Chem. Phys., **71**, (1979), 1601–1606.

[HOY] K. Hirakawa, Y. Oono and H. Yamakazi, *Experimental study on chemical turbulence. II*, Jour. Phys. Soc. Jap., **46**, (1979), 721–728.

[HGS] J.L. Hudson, K.R. Graziani and R.A. Schmitz, *Experimental evidence of chaotic states in the Belusov-Zhabotinskii reaction*, J. Chem. Phys., **67**, (1977), 3040–3044.

[ChF] J.R. Chazottes and B. Fernndez, *Dynamics of Coupled Map Lattices and of Related Spatially Extended Systems*, Lecture Notes in Physics, **671**, 2005.

[ChL] G. Chen and S. T. Liu, *On spatial periodic orbits and spatial chaos*, Int. J. of Bifur. Chaos, **13**, 935-941, 2003.

[Da] R. A. Dana and L. Montrucchio, *Dynamical Complexity in Duopoly Games*, J. Econom. Theory, **40** (1986), 40–56

[De] R.L. Devaney, *An Introduction to Chaotics Dynamical Systems*, Benjamin/Cummings, Menlo Park, CA., 1986.

[Di] E.I. Dinaburg, *A connection between various entropy characterizations of dynamical systems*, Izv. Akad. Nauk SSSR Ser. Mat., **35** (1971), 324–366

[F] H. Furnsterbeg, *Recurrence in Ergodic Theory and Combinational Number Theory.* Princeton University Press. XI, Princeton, New Jersey, 1981.

[K] K. Kaneko, *Globally Coupled Chaos Violates Law of Large Numbers*, Phys. Rev. Lett., **65**, 1391-1394, 1990.

[KO] M. Kohmoto and .Oono, *Discrete model of Chemical Turbulence*, Phys. Rev. Lett., **55**, 2927 - 2931, 1985.

[KW] K. Kaneko and H.F. Willeboordse, *Bifurcations and spatial chaos in an open flow model*, Phys. Rew. Lett., **73**, 533-536, 1994.

[Po] B. Van der Pool, *Forced oscilations in a circuit with nonlinear resistence*, London, Edinburgh and Dublin Phil. Mag, **3** (1927), 109–123.

[Pu] T. Puu, *Chaos in Duopoly Pricing*, Chaos, Solitions and Fractals, **1** (1991), 573–581.

[W] P. Walters, *An introduction to ergodic theory.* Springer, New York, 1982.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Information-Theoretic Approach to Kinetic-energy Functionals: The Nearly Uniform Electron Gas

Luca M. Ghiringhelli[1], Luigi Delle Site[1], Ricardo A. Mosna[2] and Ian P. Hamilton[3]

[1] *Max-Planck-Institute for Polymer Research, Ackermannweg 10, D 55021 Mainz Germany*

[2] *Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, C.P. 6065, 13083-859, Campinas, SP, Brazil*

[3] *Department of Chemistry, Wilfrid Laurier University, Waterloo, Canada N2L 3C5*

emails: `ghiluca@mpip-mainz.mpg.de`, `dellsite@mpip-mainz.mpg.de`, `mosna@ime.unicamp.br`, `ihamilton@wlu.ca`

Abstract

We strengthen the connection between information theory and quantum mechanical systems using a recently developed dequantization procedure which results in a decomposition of the kinetic energy as the sum of a classical term and a purely quantum term. For the nearly uniform electron gas we thereby approximate the noninteracting kinetic energy as the sum of the Thomas-Fermi term and the Weizsäcker term (which is identical in information content to the Fisher information). Electron correlation is included via a nonlocal analytical expression which is a functional of the $(N\text{-}1)$-conditional probability density. This expression is evaluated via a statistically rigorous Monte-Carlo procedure to obtain the correlation energy as a functional of the electron density. We then approximate this functional as a (local) Shannon entropy term. Thus the kinetic energy is expressed as the Thomas-Fermi term plus two terms from information theory: the Fisher information, which is a measure of electron localization, and the Shannon entropy, which is a measure of electron delocalization.

Key words: kinetic-energy functionals, Weizsäcker term, Fisher information, Shannon entropy
MSC 2000: 62B10,81Q99,00B25

*Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.*

Dual Tableau for a multimodal logic for order of magnitude qualitative reasoning with bidirectional negligibility

J. Golińska-Pilarek[1] and E. Muñoz-Velasco[2]

[1] *Institute of Philosophy, Warsaw University, Poland
National Institute of Telecommunications, Warsaw, Poland*

[2] *Dept. Applied Mathematics, University of Málaga, Spain*

emails: `j.golinska-pilarek@itl.waw.pl`, `emilio@ctima.uma.es`

Abstract

We present a relational proof system in the style of dual tableaux for the relational logic associated with a multimodal propositional logic for order of magnitude qualitative reasoning with a bidirectional relation of negligibility. We study soundness and completeness of the proof system and we show how it can be used for verification of validity of formulas of the logic on the basis of an example.

Key words: relational logics, dual tableau systems, multimodal propositional logic, order-of-magnitude qualitative reasoning

1 Introduction

The use of models to represent different scientific and engineering situations leads to qualitative reasoning as a good possibility when the traditional numeric methods are limited. Qualitative Reasoning (QR) provides an intermediate level between discrete and continuous models [20]. A form of QR is to manage numerical data in terms of orders of magnitude (see, for example, [7, 14, 17, 19]). Two approaches to order of magnitude reasoning have been identified in [21]: Absolute Order of Magnitude, which is represented by a partition of the real line $\mathbb{R}$, where each element of $\mathbb{R}$ belongs to a qualitative class and Relative Order of Magnitude, introducing a family of binary order of magnitude relations which establish different comparison relations in $\mathbb{R}$ (e.g., *comparability*, *negligibility* and *closeness*). In general, both models need to be combined to capture all the relevant information.

Several logics have been defined to use QR in different contexts, e.g. spatial and temporal reasoning [1,22]. In particular, logics dealing with order of magnitude reasoning have been developed in [3–5] by combining the absolute and relative approaches, that is, defining different qualitative relations by using the intervals provided by a specific absolute order of magnitude model.

In this paper, we focus our attention on the multimodal propositional logic $\mathcal{L}(MQ)^N$ presented in [4], which uses the absolute order of magnitude model with the real line divided in seven intervals to define a binary relation of negligibility. This negligibility relation is *bidirectional*, that is it allows us to compare positive and negative numbers. Moreover, this relation has good properties with respect to the sum and product of real numbers, which is very useful in the applications (see, for example [17]).

It is well known that one of the main advantages in the use of the logic formalism is the possibility of having automated deduction systems. For this reason, we present a relational proof system in the style of dual tableaux for the relational logic associated to the multimodal logic in question. We prove that the system enables us to verify validity of formulas of this logic. This relational system is founded on the Rasiowa-Sikorski system (RS) for the first-order logic [18] extended with the rules for equality predicate in [11]. The election of this method has many advantages [12]: a clear-cut method of generating rules of the system from the semantics, the resulting deduction system well suited for automated deduction purposes, a standard and intuitively simple way of proving completeness by constructing a counter-model for a non-provable formula out of its wrong decomposition tree in an RS system and an almost automatic way of transforming a complete RS system into a complete Gentzen calculus system.

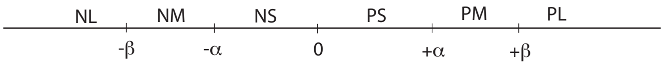
In the construction of the system, we apply the method known for various non-classical logics, (see e.g. [16]). Firstly, we construct a relational logic appropriate for the multimodal logic $\mathcal{L}(MQ)^N$. Then, we define a validity preserving translation from the language of $\mathcal{L}(MQ)^N$ to the language of the relational logic. Finally, we construct a complete and sound relational proof system for the relational logic appropriate for $\mathcal{L}(MQ)^N$. The relational logic considered in this paper is based on the classical relational logic of binary relations with relational constants 1 and $1'$, which provides a means for proving the identities valid in the class of full relation algebras (see e.g., [10, 16]). The proof system developed in the paper is the extension of the proof system for the classical relational logic originated in [15]. In constructing deduction rules of the system, we follow the general principles of defining relational deduction rules presented in [13]. Another approach to relational logics for order of magnitude reasoning has been presented in [6].

The existence of automated deduction systems give us also the possibility of implementation. In [8] there is an implementation of the proof system for the classical relational logic and in [9] an implementation of translation procedures from non-classical logics to relational logic is presented. Focusing our attention on logics for order of magnitude reasoning, in [2] a theorem prover for the system introduced in [6] has been given.

The paper is organized as follows: In Section 2, we define the syntax, semantics and the axiomatization of the logic $\mathcal{L}(MQ)^N$. In Section 3, we develop the relational logic appropriate for $\mathcal{L}(MQ)^N$ and a validity preserving translation for it. In section 4, a sound and complete relational proof system is given. Finally, in Section 5, some conclusions and future work are commented.

2 The multimodal logic $\mathcal{L}(MQ)^N$

As we have said in the introduction, we are going to work with the logic $\mathcal{L}(MQ)^N$ presented in [4]. This logic uses an absolute order of magnitude model which considers the real line $\mathbb{R}$ divided in seven equivalence classes using five landmarks, as follows:



where α, β are two positive real numbers (chosen depending on the context under consideration) such that $\alpha <_{\mathbb{R}} \beta$, being $\leq_{\mathbb{R}}$ the usual order in $\mathbb{R}$. The seven intervals are defined by NL $= (-\infty, -\beta)$, NM $= [-\beta, -\alpha)$, NS $= [-\alpha, 0)$, $[0] = \{0\}$, PS $= (0, \alpha]$, PM $= (\alpha, \beta]$ and PL $= (\beta, +\infty)$.

The labels correspond to "negative large", "negative medium", "negative small", "zero", "positive small", "positive medium", and "positive large", respectively. From this partition of the real line, we define the following negligibility relation.

Given $\alpha, \beta \in \mathbb{R}$, such that $0 <_{\mathbb{R}} \alpha <_{\mathbb{R}} \beta$, we say that x is *negligible* with respect to y, in symbols $x \, N_{\mathbb{R}} \, y$, iff, we have one of the following possibilities:

$$(i) \quad x = 0 \qquad (ii) \quad x \in \mathrm{NS} \cup \mathrm{PS} \text{ and } y \in \mathrm{NL} \cup \mathrm{PL}$$

Note that item (i) above corresponds to the intuitive idea that 0 is negligible with respect to any real number and item (ii) corresponds to the intuitive idea that a number *sufficiently small* is negligible with respect to any number *sufficiently large*, independently of the sign of these numbers. For this reason, we say that our negligibility relation is bidirectional.

Using the idea of the previous definition, we construct a logic where the five landmarks $-\beta$, $-\alpha$, 0, α, and β are replaced, respectively, by the following elements of its language: c_1, c_2, c_3, c_4, and c_5, while the negligibility relation N is defined as an accessibility relation obtained from these landmarks.

2.1 Syntax of $\mathcal{L}(MQ)^N$

We consider the language of $\mathcal{L}(MQ)^N$ as a multimodal propositional language with a family of modal operators determined by accessibility relations. Expressions of the language are constructed with symbols from the following pairwise disjoint sets:

- $\mathcal{V}$ a set of propositional variables,
- $\mathcal{C} = \{c_i \mid i \in \{1, \ldots, 5\}\}$ a set of specific constants,
- $\mathcal{A} = \{<, N\}$ a set of accessibility relational constants,

- $\{\neg, \wedge, \vee, \rightarrow, \overrightarrow{\Box}, \overleftarrow{\Box}, \Box_N, \overline{\Box}_N\}$ the set of propositional operations and the specific modal connectives. [1]

The set For of $\mathcal{L}(MQ)^N$*-formulas* is the smallest set satisfying the following conditions:

1. $\mathcal{V} \cup \mathcal{C} \subseteq For$,
2. If $\varphi, \psi \in For$, then $\neg\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \rightarrow \psi$, $\overrightarrow{\Box}\, \varphi$, $\overleftarrow{\Box}\, \varphi$, $\Box_N \varphi$ and $\overline{\Box}_N \varphi \in For$.

2.2 Semantics of $\mathcal{L}(MQ)^N$

We define the basic concepts of the semantics of our logic.

An $\mathcal{L}(MQ)^N$*-model* is a tuple $\mathcal{M} = (U, m)$, where U is a non empty set and m is a meaning function satisfying the following conditions:

1. $m(p) \subseteq U$ for $p \in \mathcal{V}$,
2. $m(<)$ is a strict linear ordering on U, that is, for all $s, s', s'' \in U$ the following conditions are satisfied:

   (Irref) $(s, s) \notin m(<)$,
   (Trans) if $(s, s') \in m(<)$ and $(s', s'') \in m(<)$, then $(s, s'') \in m(<)$,
   (Lin) $(s, s') \in m(<)$ or $(s', s) \in m(<)$ or $s = s'$,

3. $m(c_i) \in U$ for every $i \in \{1, \ldots, 5\}$, and $(m(c_i), m(c_{i+1})) \in m(<)$ for every $i \in \{1, \ldots, 4\}$,
4. $m(N) = G_1 \cup G_2 \cup G_3 \subseteq U \times U$, where:
   $G_1 = \{(s, s')\colon s = m(c_3)\}$,
   $G_2 = \{(s, s')\colon (\lambda \text{ or } \mu) \text{ and } (\gamma \text{ or } \delta) \text{ and } \zeta\}$,
   $G_3 = \{(s, s')\colon (\lambda \text{ or } \mu) \text{ and } (\gamma \text{ or } \delta) \text{ and } \eta\}$,
   where $\lambda := ((m(c_2), s) \in m(<)$, $\mu := (s = m(c_2))$, $\gamma := (((s, m(c_4)) \in m(<))$, $\delta := (s = m(c_4))$, $\zeta := ((s', m(c_1)) \in m(<))$ and $\eta := ((m(c_5), s') \in m(<))$.

Note that item 4 reflects semantically our definition of negligibility.

Let φ be an $\mathcal{L}(MQ)^N$-formula and let $\mathcal{M} = (U, m)$ be an $\mathcal{L}(MQ)^N$-model. The *satisfaction of* φ *in* $\mathcal{M}$ *by* $s \in U$, $((\mathcal{M}, s) \models \varphi$ for short), is defined as usual for propositional connectives. The definition for the modal connectives is given as follows:

- $(\mathcal{M}, s) \models \overrightarrow{\Box}\, \varphi$ iff for all $s' \in U$, $(s, s') \in m(<)$ implies $(\mathcal{M}, s') \models \varphi$,
- $(\mathcal{M}, s) \models \overleftarrow{\Box}\, \varphi$ iff for all $s' \in U$, $(s', s) \in m(<)$ implies $(\mathcal{M}, s') \models \varphi$,

[1] As usual in modal logic, we use $\overrightarrow{\Diamond}$, $\overleftarrow{\Diamond}$, $\Diamond_N$, $\overline{\Diamond}_N$ as abbreviations of $\neg \overrightarrow{\Box} \neg$, $\neg \overleftarrow{\Box} \neg$, $\neg\Box_N\neg$, and $\neg\overline{\Box}_N\neg$, respectively.

- $(\mathcal{M}, s) \models \Box_N \varphi$ iff for all $s' \in U$, $(s, s') \in m(N)$ implies $(\mathcal{M}, s') \models \varphi$,
- $(\mathcal{M}, s) \models \overline{\Box}_N \varphi$ iff for all $s' \in U$, $(s', s) \in m(N)$ implies $(\mathcal{M}, s') \models \varphi$.

We say that an $\mathcal{L}(MQ)^N$-formula φ is *satisfiable* if, and only if, there exist an $\mathcal{L}(MQ)^N$-model $\mathcal{M}$ and $s \in U$ such that $(\mathcal{M}, s) \models \varphi$. An $\mathcal{L}(MQ)^N$-formula φ is *true* in an $\mathcal{L}(MQ)^N$-model $\mathcal{M} = (U, m)$ whenever $(\mathcal{M}, s) \models \varphi$ for all $s \in U$. An $\mathcal{L}(MQ)^N$-formula φ is $\mathcal{L}(MQ)^N$*-valid*, denoted by $\models \varphi$ whenever it is true in all $\mathcal{L}(MQ)^N$-models.

2.3 Axiom system for $\mathcal{L}(MQ)^N$

The axiom system for $\mathcal{L}(MQ)^N$ consists of all the tautologies of classical propositional logic together with the following axiom schemata:

Axiom schemata for modal connectives:

K1 $\overrightarrow{\Box}\,(\varphi \to \psi) \to (\overrightarrow{\Box}\,\varphi \to \overrightarrow{\Box}\,\psi)$ **K2** $\varphi \to \overrightarrow{\Box}\overleftarrow{\Diamond}\,\varphi$ **K3** $\overrightarrow{\Box}\,\varphi \to \overrightarrow{\Box}\overrightarrow{\Box}\,\varphi$

K4 $(\overrightarrow{\Box}\,(\varphi \vee \psi) \wedge \overrightarrow{\Box}\,(\overrightarrow{\Box}\,\varphi \vee \psi) \wedge \overrightarrow{\Box}\,(\varphi \vee \overrightarrow{\Box}\,\psi)) \to (\overrightarrow{\Box}\,\varphi \vee \overrightarrow{\Box}\,\psi)$

Axiom schemata for constants: (being $i \in \{1, \ldots, 5\}$ and $j \in \{1, \ldots, 4\}$)

C1 $\overleftarrow{\Diamond}\,c_i \vee c_i \vee \overrightarrow{\Diamond}\,c_i$ **C2** $c_i \to (\overleftarrow{\Box}\,\neg c_i \wedge \overrightarrow{\Box}\,\neg c_i)$ **C3** $c_j \to \overrightarrow{\Diamond}\,c_{j+1}$

Axiom schemata for negligibility connectives:

N1 $\Box_N(\varphi \to \psi) \to (\Box_N\varphi \to \Box_N\psi)$ **N2** $\varphi \to \Box_N\overline{\Diamond}_N\varphi$
N3 $(\overleftarrow{\Box}\,\varphi \wedge \varphi \wedge \overrightarrow{\Box}\,\varphi) \to \Box_N\varphi$ **N4** $(\overrightarrow{\Diamond}\,c_2 \vee \overleftarrow{\Diamond}\,c_4) \to \Box_N(\varphi \wedge \neg\varphi)$
N5 $c_3 \to (\Box_N\varphi \to (\overleftarrow{\Box}\,\varphi \wedge \varphi \wedge \overrightarrow{\Box}\,\varphi))$
N6 $(\neg c_3 \wedge (c_2 \vee (\overleftarrow{\Diamond}\,c_2 \wedge \overrightarrow{\Diamond}\,c_4) \vee c_4)) \to \Box_N(\overrightarrow{\Diamond}\,c_1 \vee \overleftarrow{\Diamond}\,c_5)$
N7 $(\neg c_3 \wedge (c_2 \vee (\overleftarrow{\Diamond}\,c_2 \wedge \overrightarrow{\Diamond}\,c_4) \vee c_4)) \to (\Box_N\varphi \to (\overleftarrow{\Box}\,(\overrightarrow{\Diamond}\,c_1 \to \varphi) \wedge \overrightarrow{\Box}\,(\overleftarrow{\Diamond}\,c_5 \to \varphi))$

We also consider as axioms the corresponding mirror images of axioms **K1**-**K4**, and axioms **N1**-**N3**. Moreover, we consider the following Rules of Inference:

(MP) Modus Ponens (R$\overrightarrow{\Box}$) If $\vdash \varphi$ then $\vdash \overrightarrow{\Box}\,\varphi$ (R$\overleftarrow{\Box}$) If $\vdash \varphi$ then $\vdash \overleftarrow{\Box}\,\varphi$

3 The relational logic $\mathcal{R}(MQ)^N$

In this section we present the syntax and semantics of the relational logic appropriate for the multimodal logic $\mathcal{L}(MQ)^N$.

3.1 Syntax of $\mathcal{R}(MQ)^N$

The vocabulary of the language of $\mathcal{R}(MQ)^N$ consists of the pairwise disjoint sets enumerated below:

- A (nonempty) set $\mathbb{OV} = \{x, y, z, \ldots\}$ of object variables,

- A set $\mathbb{OC} = \{c_1, \ldots, c_5\}$ of object constants,
- A (nonempty) set $\mathbb{RV} = \{R, S, \ldots\}$ of binary relational variables,
- A set $\mathbb{RC} = \{1, 1', <, N\} \cup \{\Psi_1, \ldots, \Psi_5\}$ of relational constants,
- A set $\mathbb{OP} = \{-, \cup, \cap, ;, ^{-1}\}$ of relational operation symbols.

The set of *relational terms* $\mathbb{RT}$ is the smallest set of expressions that includes $\mathbb{RA} = \mathbb{RV} \cup \mathbb{RC}$ and is closed with respect to the operation symbols from $\mathbb{OP}$. The set $\mathbb{FR}$ of $\mathcal{R}(MQ)^N$*-formulas* (or, simply formulas if it is clear from the context), consists of expressions of the form xRy, where $x, y \in \mathbb{OS} = \mathbb{OV} \cup \mathbb{OC}$ and $R \in \mathbb{RT}$. R is said to be an *atomic relational term* whenever $R \in \mathbb{RA}$. xRy is said to be an *atomic formula* whenever R is an atomic relational term.

We will use the following notation, very useful in the rest of the paper:

$$\begin{array}{llll} n_1(x,y) := c_3 1' x & n_2(x,y) := c_4 1' x & n_3(x,y) := c_2 1' x & n_4(x,y) := c_5 < y \\ n_5(x,y) := y < c_1 & n_6(x,y) := c_2 < x & n_7(x,y) := x < c_4 & \end{array}$$

If $\varphi(x,y)$ is a formula xRy, then by $-\varphi(x,y)$ we denote the formula $x{-}Ry$.

Finally, we define the following sequences of formulas:

$$\begin{array}{ll} K_1(x,y) := n_1(x,y) & K_2(x,y) := n_2(x,y), n_5(x,y) \\ K_3(x,y) := n_2(x,y), n_4(x,y) & K_4(x,y) := n_3(x,y), n_5(x,y) \\ K_5(x,y) := n_3(x,y), n_4(x,y) & K_6(x,y) := n_6(x,y), n_7(x,y), n_5(x,y) \\ K_7(x,y) := n_6(x,y), n_7(x,y), n_4(x,y) & \end{array}$$

$$\begin{array}{l} H_1(x,y) := n_1(x,y), n_2(x,y), n_3(x,y), n_6(x,y) \\ H_2(x,y) := n_1(x,y), n_2(x,y), n_3(x,y), n_7(x,y) \\ H_3(x,y) := n_1(x,y), n_4(x,y), n_5(x,y) \end{array}$$

$$-K_l(x,y) := (-n_k(x,y))_{n_k(x,y) \in K_l(x,y)}, \text{ for } l \in \{1, \ldots, 7\}$$

3.2 Semantics of $\mathcal{R}(MQ)^N$

Now, we present the main definitions in the semantics of the relational logic.

An $\mathcal{R}(MQ)^N$*-model* is a pair $\mathcal{M}' = (U', m')$, where U' is a non-empty set and m' a meaning function, $m' : \mathbb{RA} \cup \mathbb{OC} \to \mathcal{P}(U' \times U') \cup U'$, defined as follows:

1. $m'(1')$ is an equivalence relation on U',
2. $m'(1'); m'(R) = m'(R); m'(1') = m'(R)$ for every $R \in \mathbb{RA}$ (extensionality prop.),
3. $m'(1) = U' \times U'$,

4. $m'(<)$ is a relation on U', which satisfies, for all $s, s', s'' \in U'$:

   (Irref) $(s,s) \notin m'(<)$
   (Trans) if $(s,s') \in m'(<)$ and $(s',s'') \in m'(<)$, then $(s,s'') \in m'(<)$
   (Lin) $(s,s') \in m'(<)$ or $(s',s) \in m'(<)$ or $(s,s') \in m'(1')$,

5. $m'(c_i) \in U'$ and $(m'(c_i), m'(c_{i+1})) \in m'(<)$, for $i \in \{1,\ldots,5\}$,

6. $m'(\Psi_i) = \{(s,s') \in U' \times U' : (s, m'(c_i)) \in m'(1')\}$,

7. $(s,s') \in m'(N)$ iff $\displaystyle\bigwedge_{m\in\{1,2,3\}} H^*_m(s,s')$ iff $\displaystyle\bigvee_{l\in\{1,\ldots,7\}} K^*_l(s,s')$, where:

   $n^*_1(s,s') := (s, m'(c_3)) \in m'(1')$ $\quad n^*_2(s,s') := (s, m'(c_4)) \in m'(1')$
   $n^*_3(s,s') := (s, m'(c_2)) \in m'(1')$ $\quad n^*_4(s,s') := (m'(c_5), s') \in m'(<)$
   $n^*_5(s,s') := (s', m'(c_1)) \in m'(<)$ $\quad n^*_6(s,s') := (m(c_2), s) \in m'(<)$
   $n^*_7(s,s') := (s, m'(c_4)) \in m'(<)$

   $$H^*_m(s,s') := \bigvee_{\{h \mid n_h(x,y) \in H_m(x,y)\}} n^*_h(s,s') \text{ for } m \in \{1,2,3\}$$
   $$K^*_l(s,s') := \bigwedge_{\{k \mid n_k(x,y) \in K_l(x,y)\}} n^*_k(s,s') \text{ for } l \in \{1,\ldots,7\},$$

   where $\bigwedge$ and $\bigvee$ are used here as meta-connectives.

8. m' extends to all compound relational terms as usual, that is:

   $m'(-R) = m(1) \cap -m'(R) \quad m'(R^{-1}) = m'(R)^{-1} \quad m'(R;S) = m'(R); m'(S)$
   $m'(R \cap S) = m'(R) \cap m'(S) \quad m'(R \cup S) = m'(R) \cup m'(S)$

An $\mathcal{R}(MQ)^N$-model $\mathcal{M}' = (U', m')$ is said to be *standard* whenever $m'(1')$ is the identity on U', that is $m'(1') = \{(x,x) : x \in U'\}$ [2]. The class of standard models is denoted by $\mathcal{R}^*(MQ)^N$ and we use in this paper the term *standard model* or $\mathcal{R}^*(MQ)^N$*-model* indistinctly. An $\mathcal{R}(MQ)^N$*-valuation* in an $\mathcal{R}(MQ)^N$-model $\mathcal{M}' = (U', m')$ is a function $v : \mathbb{OS} \to U'$ such that $v(c_i) = m'(c_i)$, for every $i \in \{1,\ldots,5\}$. Let xRy be an $\mathcal{R}(MQ)^N$-formula and let $\mathcal{M}' = (U', m')$ be an $\mathcal{R}(MQ)^N$-model. A formula xRy is said to be *satisfied in* $\mathcal{M}'$ *by* v ($\mathcal{M}', v \models xRy$ for short) whenever $(v(x), v(y)) \in m'(R)$. A formula xRy is *true* in $\mathcal{M}'$ if it is satisfied in $\mathcal{M}'$ by all valuations v. xRy is said to be $\mathcal{R}(MQ)^N$*-valid*, if it is true in all $\mathcal{R}(MQ)^N$-models. Moreover, a formula is said to be $\mathcal{R}^*(MQ)^N$*-valid* whenever it is true in all standard models.

As we will see in Section 4, condition 7 reflects the definition of the negligibility relation in a suitable form.

Now, we develop the validity preserving *translation function* $t : For \to \mathbb{RT}$ assigning relational terms to modal formulas. We start with an assignment t' of relational variables to all propositional variables, $t'(p) = R_p$ where $R_p \in \mathbb{RV}$. Then we define:

[2] Note that in standard models $m'(<)$ is a strict linear ordering on U'.

$t(p) = t'(p); 1$, for every propositional variable $p \in \mathcal{V}$,

$t(c_i) = \Psi_i; 1$, for every $i \in \{1, \ldots, 5\}$,

t extends to all compound $\mathcal{L}(MQ)^N$-formulas as follows:

$$\begin{array}{lll}
t(\neg\varphi) = -t(\varphi) & t(\varphi \vee \psi) = t(\varphi) \cup t(\psi) & t(\varphi \wedge \psi) = t(\varphi) \cap t(\psi) \\
t(\varphi \to \psi) = -t(\varphi) \cup t(\psi) & t(\overrightarrow{\Box}\,\varphi) = -(<; -t(\varphi)) & t(\overleftarrow{\Box}\,\varphi) = -(<^{-1}; -t(\varphi)) \\
t(\Box_N\varphi) = -(N; -t(\varphi)) & t(\overline{\Box}_N\varphi) = -(N^{-1}; -t(\varphi)) &
\end{array}$$

Now, we want to obtain the semantic relationship between the multimodal and relational logics. To begin with, we associate (mutually) $\mathcal{L}(MQ)^N$-models and $\mathcal{R}^*(MQ)^N$-models using the translation function.

Proposition 1 *For every $\mathcal{L}(MQ)^N$-model $\mathcal{M} = (U, m)$ and for every $\mathcal{L}(MQ)^N$-formula ψ there is a standard $\mathcal{R}(MQ)^N$-model $\mathcal{M}'$ such that, for all object variables x and y, ψ is true in $\mathcal{M}$ iff $x\, t(\psi)\, y$ is true in $\mathcal{M}'$.*

Proposition 2 *For every $\mathcal{R}^*(MQ)^N$-model $\mathcal{M}' = (U', m')$ and for every $\mathcal{L}(MQ)^N$-formula ψ there is an $\mathcal{L}(MQ)^N$-model $\mathcal{M}$ such that, for all object variables x and y, ψ is true in $\mathcal{M}$ iff $x\, t(\psi)\, y$ is true in $\mathcal{M}'$.*

From Propositions 1 and 2 we obtain the desired result:

Theorem 3 *For every $\mathcal{L}(MQ)^N$-formula ψ and for all object variables x and y, we have that ψ is $\mathcal{L}(MQ)^N$-valid iff $xt(\psi)y)$ is $\mathcal{R}^*(MQ)^N$-valid.*

4 Relational proof system for $\mathcal{R}(MQ)^N$

The proof system for logic $\mathcal{R}(MQ)^N$ presented in this section belongs to the family of dual tableau systems. Dual tableau systems are determined by axiomatic sets of formulas and rules which apply to finite sets of formulas. The axiomatic sets take the place of axioms. There are two groups of rules: the *decomposition rules*, which reflect definitions of the standard relational operations and the *specific rules* which reflect the properties of the specific relations assumed in $\mathcal{R}(MQ)^N$-models. The rules have the following general form:

$$(*) \qquad \frac{\Phi}{\Phi_1 \,|\, \ldots \,|\, \Phi_n}$$

where $\Phi_1, \ldots, \Phi_n$ are finite non-empty sets of formulas, $n \geq 1$, and Φ is a finite (possibly empty) set of formulas. Φ is called the *premise* of the rule, and $\Phi_1, \ldots, \Phi_n$ are called its *conclusions*. A rule of the form $(*)$ is said to be *applicable* to a set X of formulas whenever $\Phi \subseteq X$. As a result of application of a rule of the form $(*)$ to a set X, we obtain the sets $(X \setminus \Phi) \cup \Phi_i$, $i = 1, \ldots, n$. As usual, any concrete rule will always be presented in a short form, that is we will omit set brackets. We say that an object variable in a rule is *new* whenever it appears in a conclusion of the rule and does not appear in its premise.

Let $x, y, \in \mathbb{OS}$ and $R, S \in \mathbb{RT}$. *Decomposition rules* of the system have the following forms, for any object symbol z and for a new object variable w:

$$(\cup)\ \frac{x(R\cup S)y}{xRy, xSy} \qquad (-\cup)\ \frac{x-(R\cup S)y}{x-Ry \mid x-Sy} \qquad (;)\ \frac{x(R;S)y}{xRz, x(R;S)y|zSy, x(R;S)y}$$

$$(\cap)\ \frac{x(R\cap S)y}{xRy|xSy} \qquad (-\cap)\ \frac{x-(R\cap S)y}{x-Ry, x-Sy} \qquad (-;)\ \frac{x-(R;S)y}{x-Rw, w-Sy}$$

$$(-)\ \frac{x--Ry}{xRy} \qquad (^{-1})\ \frac{xR^{-1}y}{yRx} \qquad (-^{-1})\ \frac{x-R^{-1}y}{y-Rx}$$

Let $x, y, z \in \mathbb{OS}$, $R \in \mathbb{RA}$ and $i \in \{1, \ldots, 5\}$. *Specific rules* have the following forms, for any object symbol z:

$$(1'1)\ \frac{xRy}{xRz, xRy|y1'z, xRy} \qquad (1'2)\ \frac{xRy}{x1'z, xRy|zRy, xRy}$$

$$(\text{Irref}<)\ \frac{}{x<x} \qquad (\text{Tran}<)\ \frac{x<y}{x<y, x<z|x<y, z<y}$$

$$(C_i1)\ \frac{}{x\Psi_i y \mid x-\Psi_i y} \qquad (C_i2)\ \frac{x\Psi_i y}{x\Psi_i y, x1'c_i}$$

$$(C_i3)\ \frac{x-\Psi_i y}{x-\Psi_i y, x-1'c_i} \qquad (\text{N1})\ \frac{}{xNy \mid x-Ny}$$

$$(\text{N2})\ \frac{xNy}{xNy, H_1(x,y)|xNy, H_2(x,y)|xNy, H_3(x,y)} \qquad (\text{N3})\ \frac{x-Ny}{x-Ny, -K_1(x,y)|\ldots|x-Ny, -K_7(x,y)}$$

where $H_m(x, y)$ and $K_l(x, y)$, $m \in \{1, 2, 3\}$, $l \in \{1, \ldots, 7\}$, are defined previously.

A finite set of formulas is $\mathcal{R}(MQ)^N$-*axiomatic* whenever it includes either of the following subsets, for all $x, y \in \mathbb{OS}$, $i \in \{1, \ldots, 4\}$, and for every $R \in \mathbb{RT}$:

$$(\text{Ax1})\ \{x1'x\} \quad (\text{Ax2})\ \{x1y\} \quad (\text{Ax3})\ \{xRy, x-Ry\} \quad (\text{Ax4})\ \{c_i < c_{i+1}\} \quad (\text{Ax5})\ \{x<y, y<x, x1'y\}$$

Now, we give some definitions and results needed to obtain the soundness of the system.

A finite set of $\mathcal{R}(MQ)^N$-formulas $\{xR_1y, \ldots, xR_2y\}$ is said to be an $\mathcal{R}(MQ)^N$-*set* whenever for every $\mathcal{R}(MQ)^N$-model $\mathcal{M}'$ and for every valuation v in $\mathcal{M}'$ there exists $i \in \{1, \ldots, n\}$ such that $\mathcal{M}', v \models \varphi_i$. A rule $\frac{\Phi}{\Phi_1|\ldots|\Phi_n}$ is $\mathcal{R}(MQ)^N$-*correct* whenever for every finite set X of $\mathcal{R}(MQ)^N$-formulas $X \cup \Phi$ is an $\mathcal{R}(MQ)^N$-set iff $X \cup \Phi_j$ is an $\mathcal{R}(MQ)^N$-sets for every $j \in \{1, \ldots, n\}$. Due to the semantics we obtain the following:

Proposition 4

1. *The decomposition rules are $\mathcal{R}(MQ)^N$-correct.*
2. *The specific rules are $\mathcal{R}(MQ)^N$-correct.*
3. *The axiomatic sets are $\mathcal{R}(MQ)^N$-sets.*

We define an $\mathcal{R}(MQ)^N$-*proof tree* for xRy as a tree with the following properties:

- the formula xRy is at the root of this tree,

- each node except the root is obtained by an application of an $\mathcal{R}(MQ)^N$-rule to its predecessor node,
- a node does not have successors whenever it is an $\mathcal{R}(MQ)^N$-axiomatic set.

Due to the forms of the rules for atomic formulas, we can say that if a node of an $\mathcal{R}(MQ)^N$-proof tree does not contain an $\mathcal{R}(MQ)^N$-axiomatic subset and contains an $\mathcal{R}(MQ)^N$-formula xRy or $x{-}Ry$, for atomic R, then all of its successors contain this formula as well.

A branch of an $\mathcal{R}(MQ)^N$-proof tree is said to be *closed* whenever it contains a node with an $\mathcal{R}(MQ)^N$-axiomatic set of formulas. A *closed tree* is an $\mathcal{R}(MQ)^N$-proof tree such that all of its branches are closed. A formula xRy is *$\mathcal{R}(MQ)^N$-provable* whenever there is a closed proof tree for xRy.

From Proposition 4, we obtain the soundness of the system:

Theorem 5 (Soundness) *Let xRy be an $\mathcal{R}(MQ)^N$-formula. If xRy is $\mathcal{R}(MQ)^N$-provable, then it is $\mathcal{R}(MQ)^N$-valid.*

Since $\mathcal{R}(MQ)^N$-validity implies $\mathcal{R}^*(MQ)^N$-validity, we obtain the following:

Corolary 6 *If xRy is $\mathcal{R}(MQ)^N$-provable, then it is $\mathcal{R}^*(MQ)^N$-valid.*

As usual in the proof theory a concept of completeness of a non-closed proof tree is needed. Intuitively, completeness of a non-closed tree means that all the rules that can be applied have been applied. By abusing the notation, for any branch b and for any set of formulas X, by $X \in b$ (resp. $X \notin b$) we mean that every formula from X belongs to b (resp. does not belong to b).

Completion Conditions
A non-closed branch b of a proof tree is said to be *complete* whenever for all $x, y \in \mathbb{OS}$ it satisfies the following completion conditions:

Cpl($\cup$) (resp. Cpl($-\cap$)) If $x(R \cup S)y \in b$ (resp. $x{-}(R \cap S)y \in b$), then both $xRy \in b$ (resp. $x{-}Ry \in b$) and $xSy \in b$ (resp. $x{-}Sy \in b$).
Cpl($\cap$) (resp. Cpl($-\cup$)) If $x(R \cap S)y \in b$ (resp. $x{-}(R \cup S)y \in b$), then either $xRy \in b$ (resp. $x{-}Ry \in b$) or $xSy \in b$ (resp. $x{-}Sy \in b$).
Cpl($-$) If $x(--R)y \in b$, then $xRy \in b$.
Cpl($^{-1}$) If $xR^{-1}y \in b$, then $yRx \in b$.
Cpl($-^{-1}$) If $x{-}R^{-1}y \in b$, then $y{-}Rx \in b$.
Cpl(;) If $x(R;S)y \in b$, then for every $z \in \mathbb{OS}$, either $xRz \in b$ or $zSy \in b$.
Cpl($-$;) If $x{-}(R;S)y \in b$, then for some $w \in \mathbb{OV}$, both $x{-}Rw \in b$ and $w{-}Sy \in b$.
Cpl($1'1$) If $xRy \in b$ for some $R \in \mathbb{RA}$, then for every $z \in \mathbb{OS}$, either $xRz \in b$ or $y1'z \in b$.
Cpl($1'2$) If $xRy \in b$ for some $R \in \mathbb{RA}$, then for every $z \in \mathbb{OS}$, either $x1'z \in b$ or $zRy \in b$.
Cpl(C_i1) Either $x\Psi_i y \in b$ or $x{-}\Psi_i y \in b$.
Cpl(C_i2) If $x\Psi_i y \in b$ then $x1'c_i \in b$.
Cpl(C_i3) If $x{-}\Psi_i y \in b$ then $x{-}1'c_i \in b$.

Cpl(Irref<) For every $x \in \mathbb{OS}$, $x < x \in b$.
Cpl(Tran<) If $x < y \in b$, then for every $z \in \mathbb{OS}$, either $x < z \in b$ or $z < y \in b$.
Cpl($N1$) Either $xNy \in b$ or $x{-}Ny \in b$.
Cpl($N2$) If $xNy \in b$ then there exists $m \in \{1, 2, 3\}$ such that $H_m(x, y) \in b$.
Cpl($N3$) If $x{-}Ny \in b$ then there exists $l \in \{1, \dots, 7\}$ such that $-K_l(x, y) \in b$.

An $\mathcal{R}(MQ)^N$-proof tree is said to be *complete* iff all of its non-closed branches are complete. A complete non-closed branch is said to be *open.* Since the set containing a subset $\{xRy, x{-}Ry\}$ is an $\mathcal{R}(MQ)^N$-axiomatic set, the following can be proved by induction:

Proposition 7 *Let b be an open branch of an $\mathcal{R}(MQ)^N$-proof tree. Then there is no $\mathcal{R}(MQ)^N$-formula xRy such that $xRy \in b$ and $x{-}Ry \in b$.*

As said in the introduction of the paper, we have a standard and intuitively simple way of proving completeness by constructing a counter-model for a non-provable formula out of its wrong decomposition tree. For this purpose, we give the following definition.

Let b be an open branch of an $\mathcal{R}(MQ)^N$-proof tree. A *branch structure* $\mathcal{M}^b$ is a pair $\mathcal{M}^b = (U^b, m^b)$, such that:

- $U^b = \mathbb{OS}$,
- $m^b(c_i) = c_i$, for every $i \in \{1, \dots, 5\}$,
- $m^b(R) = \{(x, y) \in U^b \times U^b : xRy \notin b\}$, for every $R \in \mathbb{RA}$,
- m^b extends to all compound relational terms as in $\mathcal{R}(MQ)^N$-models.

Directly by the above definition, we obtain the following:

Proposition 8 *For every open branch b, a branch structure $\mathcal{M}^b = (U^b, m^b)$ is an $\mathcal{R}(MQ)^N$-model.*

Let v^b be a valuation in a branch structure $\mathcal{M}^b$ defined as: $v^b(x) = x$ for every $x \in \mathbb{OS}$. By an easy induction, using the completion conditions and Proposition 7, we can prove the following:

Proposition 9 *Let $\mathcal{M}^b = (U^b, m^b)$ be a branch structure and let xRy be an $\mathcal{R}(MQ)^N$-formula. Then, we have:*

$$(*) \;\; \textit{If } \mathcal{M}^b, v^b \models xRy, \textit{ then } xRy \notin b.$$

Let $\mathcal{M}^b = (U^b, m^b)$ a branch structure, since $m^b(1')$ is an equivalence relation on U^b, we may define a standard $\mathcal{R}(MQ)^N$-model from $\mathcal{M}^b$.

Given $\mathcal{M}^b = (U^b, m^b)$ a branch structure, the *quotient model* $\mathcal{M}^b_q = (U^b_q, m^b_q)$ is defined as follows:

- $U^b_q = \{\|x\| : x \in U^b\}$, where $\|x\|$ is an equivalence class of $m^b(1')$ generated by x,
- $m^b_q(c_i) = \|c_i\|$, for $i \in \{1, \ldots, 5\}$,
- $m^b_q(R) = \{(\|x\|, \|y\|)) \in U^b_q \times U^b_q : (x, y) \in m^b(R)\}$, for every $R \in \mathbb{RA}$,
- m^b_q extends to all compound relational terms as in $\mathcal{R}(MQ)^N$-models.

Since $m^b(1')$ is an equivalence relation satisfying the extensionality property, the definition of $m^b_q(R)$ is correct, that is, the following condition is satisfied:

$$\text{If } (x, y) \in m^b(R) \text{ and } (x, z), (y, t) \in m^b(1'), \text{ then } (z, t) \in m^b(R).$$

By the definition of the quotient model and Proposition 8, we obtain:

Proposition 10 *The quotient model $\mathcal{M}^b_q = (U^b_q, m^b_q)$ is a standard $\mathcal{R}(MQ)^N$-model.*

Let v^b_q be a valuation in $\mathcal{M}^b_q$ such that $v^b_q(x) = \|x\|$, for every $x \in \mathbb{OS}$. By an easy induction on the complexity of formulas, the following can be proved:

Proposition 11 *Let xRy be an $\mathcal{R}(MQ)^N$-formula, then the following holds:*

$$\mathcal{M}^b, v^b \models xRy \textit{ iff } \mathcal{M}^b_q, v^b_q \models xRy$$

The above propositions enable us to prove the completeness of a relational proof system for $\mathcal{R}(MQ)^N$.

Theorem 12 (Completeness) *Let xRy be an $\mathcal{R}(MQ)^N$-formula. If xRy is $\mathcal{R}^*(MQ)^N$-valid, then it is $\mathcal{R}(MQ)^N$-provable.*

By Theorems 5 and 12 and Corollary 6 we obtain the following main theorem:

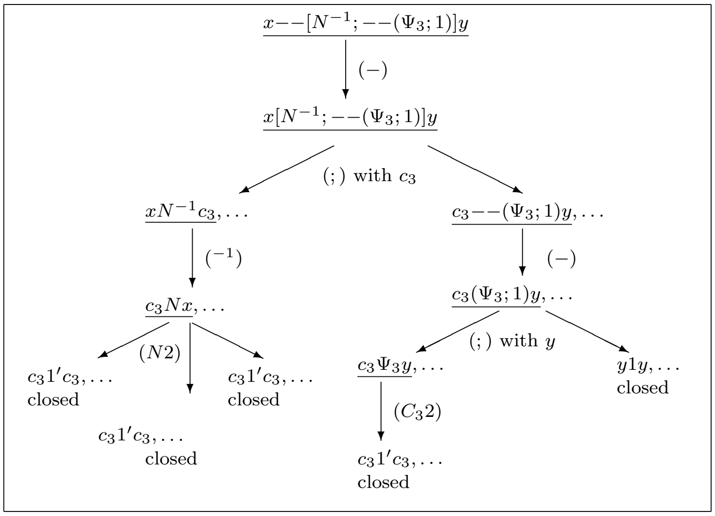
Theorem 13 (Soundness and Completeness of $\mathcal{R}(MQ)^N$)
Let xRy be an $\mathcal{R}(MQ)^N$-formula. Then xRy is $\mathcal{R}(MQ)^N$-provable iff xRy is $\mathcal{R}(MQ)^N$-valid iff xRy is $\mathcal{R}^*(MQ)^N$-valid.

Finally, by Theorems 3 and 13, we obtain the following result needed for verification of validity:

Theorem 14 (Relational Soundness and Completeness of $\mathcal{L}(MQ)^N$) *For every $\mathcal{L}(MQ)^N$-formula ψ and for all object variables x and y, the following holds:*

$$\psi \textit{ is } \mathcal{L}(MQ)^N\textit{-valid iff } xt(\psi)y \textit{ is } \mathcal{R}(MQ)^N\textit{-provable}$$

We finish this section with an example of validity checking. Consider the formula $\varphi := \overline{\Diamond}_N c_3$. It is easy to check that this formula is $\mathcal{L}(MQ)^N$-valid. The translation of φ to the relational term is $t(\varphi) = --(N^{-1}; --(\Psi_3; 1))$. The following picture presents a closed $\mathcal{R}(MQ)^N$-proof tree. It shows $\mathcal{R}(MQ)^N$-provability of the relational formula $x\tau(\varphi)y$, and by Theorem 14, it proves $\mathcal{L}(MQ)^N$-validity of φ. In each node of the proof tree, we underline the formula to which a rule has been applied.


$\underline{x--[N^{-1}; --(\Psi_3; 1)]y}$
$(-)$
$\underline{x[N^{-1}; --(\Psi_3; 1)]y}$
$(;)$ with c_3
$\underline{xN^{-1}c_3}, \dots$
$\underline{c_3--(\Psi_3; 1)y}, \dots$
$(^{-1})$
$(-)$
$\underline{c_3Nx}, \dots$
$\underline{c_3(\Psi_3; 1)y}, \dots$
$(N2)$
$(;)$ with y
$c_3 1' c_3, \dots$ closed
$c_3 1' c_3, \dots$ closed
$c_3 1' c_3, \dots$ closed
$\underline{c_3\Psi_3 y}, \dots$
$y1y, \dots$ closed
$(C_3 2)$
$c_3 1' c_3, \dots$ closed

5 Conclusions and future work

In this paper, we have introduced a relational proof system in the style of dual tableaux for the relational logic associated with the multimodal propositional logic for order of magnitude qualitative reasoning $\mathcal{L}(MQ)^N$ and we have proved its soundness and completeness. Moreover, we have shown how the proof system can be used for verification of validity and we have given an example of the relational proof of validity for a specific formula. As a future work, our plan is to adapt the implementation [2] to the specific relational system presented in the paper. On the other hand, we are planning to investigate the decidability of the logic $\mathcal{L}(MQ)^N$ and, if the answer is positive, to find a decision procedure, that is, a complete and sound relational proof system such that all of its proof trees are finite.

Acknowledgements

We thank Prof. Ewa Orłowska for suggesting the subject of the paper and for her helpful remarks on the first version of this article. The first author of the paper is a recipient of the 2007 and 2008 Grant for Young Scientists of the Foundation for Polish Science and is partially supported by the Polish Ministry of Science and Higher Education grant N N206 399134. The second author of the paper is partially supported by the Spanish research projects TIN2006-15455-C03-01 and P6-FQM-02049.

References

[1] Bennett, B., Cohn, A.G., Wolter, F. and Zakharyaschev, M. *Multi-Dimensional Modal Logic as a Framework for Spatio-Temporal Reasoning.*, Applied Intelligence **17**(3) (2002), 239–251.

[2] Burrieza, A., Mora, A., Ojeda-Aciego, M., and Orłowska, E. *Implementing a relational system for order-of-magnitude reasoning.* Technical Report (2008).

[3] Burrieza, A., Muñoz-Velasco, E., and Ojeda-Aciego, M. *A Logic for Order of Magnitude Reasoning with Negligibility, Non-closeness and Distance*, Lecture Notes in Artifical Intelligence **4788** (2007), 210-219.

[4] Burrieza, A., Muñoz-Velasco, E., and Ojeda-Aciego, M. *Order of magnitude reasoning with bidirectional negligibility.* Lect. Notes in Artificial Intelligence, **4177** (2006), 370–378.

[5] Burrieza, A. and Ojeda-Aciego, M. *A multimodal logic approach to order of magnitude qualitative reasoning with comparability and negligibility relations.* Fundamenta Informaticae **68** (2005) 21–46.

[6] Burrieza, A., Ojeda-Aciego, M., and Orłowska, E. *Relational approach to order of magnitude reasoning.* Lecture Notes in Computer Science **4342** (2006) 105-124.

[7] Dague, P. *Symbolic reasoning with relative orders of magnitude.*, Proc. 13th Intl. Joint Conference on Artificial Intelligence, Morgan Kaufmann (1993), 1509–1515.

[8] Dallien, J. and MacCaull, W. *RelDT: A relational dual tableaux automated theorem prover.*, `http://www.logic.stfx.ca/reldt/`.

[9] Formisano, A. and Orłowska, E. and Omodeo, E., *A PROLOG tool for relational translation of modal logics: A front-end for relational proof systems.*, TABLEAUX 2005 Position Papers and Tutorial Descriptions, B. Beckert (ed.), Fachberichte Informatik No 12, Universitaet Koblenz-Landau (2005) 1–10.

[10] Golińska-Pilarek, J., and Orłowska, E., *Relational logics and their applications.* Lecture Notes in Artificial Intelligence **4342** (2006) 125–161.

[11] Golińska-Pilarek, J., and Orłowska, E., *Tableaux and Dual Tableaux: Transformation of proofs.*, Studia Logica **85** (2007), 291–310.

[12] Konikowska, B., *Rasiowa-Sikorski deduction systems in computer science applications.*, Theoretical Computer Science **286** (2002), 323–366.

[13] MacCaull W., and Orłowska, E. *Correspondence results for relational proof systems with application to the Lambek calculus*, Studia Logica **71** (2002), 279–304.

[14] Mavrovouniotis, M.L., and Stephanopoulos, G., *Reasoning with orders of magnitude and approximate relations.*, Proc. 6th National Conference on Artificial Intelligence, The AAAI Press/The MIT Press (1987).

[15] Orłowska, E., *Relational interpretation of modal logics.*, in: H. Andréka, D. Monk and I. Nemeti (eds.), Algebraic Logic, Col. Math. Soc. J. Bolyai 54, North Holland, Amsterdam (1988), 443–471.

[16] Orłowska, E., *Relational proof systems for modal logics.*, in: Wansing, H. (ed.), Proof Theory of Modal Logics, Kluwer (1996), 55–77.

[17] Raiman, O. *Order of magnitude reasoning*, Artificial Intelligence **51** (1991) 11–38.

[18] Rasiowa, H., and Sikorski, R., *The Mathematics of Metamathematics.*, Polish Scientific Publishers, Warsaw 1963.

[19] Sánchez, M., Prats, F., and Piera, N., *Una formalizacin de relaciones de comparabilidad en modelos cualitativos.*, Boletn de la AEPIA (Bulletin of the Spanish Association for AI) 6 (1996), 15–22.

[20] Travé-Massuyès, L., Ironi, L. and Dague P., *Mathematical Foundations of Qualitative Reasoning.*, AI Magazine, American Asociation for Artificial Intelligence (2003), 91–106.

[21] Travé-Massuyès, L., Prats, F., Sánchez, M. and Agell, N., *Relative and absolute order-of-magnitude models unified.* Annals of Mathematics and Artificial Intelligence **45** (2005), 323–341.

[22] Wolter, F. and Zakharyaschev, M., *Qualitative spatio-temporal representation and reasoning: a computational perspective.*, in G. Lakemeyer and B. Nebel (eds.), Exploring Artificial Intelligence in the New Millenium, Morgan Kaufmann (2002).

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

New Teaching Resources to adapt Mathematics to the new European Area of Higher Education

Ascensión Hernández Encinas[1], **Luis Hernández Encinas**[2] **and Araceli Queiruga Dios**[1]

[1] *Dpt. Applied Mathematics, University of Salamanca*

[2] *Dpt. Information Processing and Coding, Applied Physics Institute, CSIC*

emails: ascen@usal.es, luis@iec.csic.es, queirugadios@usal.es

Abstract

In this study a new experience for teaching mathematics to engineering students, is presented. The aim is to provide the students a better understanding of maths algorithms and basic concepts by using specific software, in a graduate-level course. In this paper we discuss how to structure, define, and implement a web-based course as a part of the traditional classes, according to the convergence of the European Higher Education Project. The proposed course facilitates the use of new Information and Communication Technologies.

Key words: Implementation, Information and Communication Technologies, European Area of Higher Education.

1 Introduction

The creation of an European Area of Higher Education (EAHE) was proposed by the Bologna declaration in 1999, to unify university studies in Europe. The declaration emphasizes the creation of the European Area of Higher Education as a key to promote citizens' mobility and employability and the Continent's overall development [2]. Spain is one of the 46 countries involved in the Bologna Process. The corner stones of such an open space are mutual recognition of degrees and other higher education qualifications, transparency (readable and comparable degrees organised in a three-cycle structure), and European cooperation in quality assessment. University studies must be adapted to the international European context and technology development, facilitating new strategies of communication. This new situation forces Universities to renew some situations that until now seemed stable as teaching methodologies, and change their degrees and studies programmes. The use of Information and Communication Technologies (ICT) become more and more important in the higher education process, and it is considered a pre-requisite for the adaptation to the EAHE, claiming new spaces and conditions of learning, and new professional roles for lecturers [3].

Linear Algebra is a branch of Mathematics, and, in general, it is a part of the curriculum in the first course of Industrial Engineering students. The algebra course teached at the Escuela in the University of Salamanca is divided into 6 modules, concerned with the study of vector spaces, invertible linear maps or matrices, determinants or representation of a matrix in terms of its eigenvalues and eigenvectors, systems of linear equations, and a brief introduction to linear constant-coefficient systems of differential equations. In this paper, we present some educational tools to learn about Linear Algebra with Moodle environment. Moodle is an open source package, designed under pedagogical principles, in order to help educators to create effective online learning communities.

The rest of the paper is organized as follows: In section 2, we will comment the changes that are happening in the Spanish Universities to reach the European Area of Higher Education. In section 3 we will present the Moodle tools used in the University of Salamanca (`http://www.usal.es`) and how to combine them with traditional algebra classes, and finally, the conclusions will be shown in section 4.

2 Changes in higher education

The knowledge society depends for its growth on the production of new knowledge, its transmission through education and training, and its dissemination through Information and Communication Technologies [1]. As it was mentioned in the Introduction, one of the means to get the convergence of European higher education and the common goal of the Bologna Declaration is the use of the ICT in higher education. Universities face an imperative necessity to adapt and adjust to a whole series of profound changes, including increased demand, internationalisation and links with business.

Online education also refers to learning methods that, at least, partly utilize the ICT available through the Internet. What we propose to the students is to use the online methods to get a more complete education in specific subjects. The online education is a new method of education, very different from traditional education, that take advantage of new media, new ways to communicate, and the design of new educational experiences. Educators are thus utilizing the Internet for professional networking, regionally and globally, they learn from one another about the new media and their applications to education [7], and renew their knowledge in virtually fields of enquiry.

ICT have changed from being considered as a mere object of use towards an instrument of support in the educational innovation [6]. They affect to different aspects in relation to traditional education, as the change in the role of the teacher, who has changed from a simple transmitter of knowledge to be a mediator in the construction of the knowledge of the students; the role of the student has changed as the traditional educative models do not adjust to the processes of learning by means of the use of the ICT [5]. Finally, it is important to take into account that the use of new technologies does not require the invention of new methodologies, but it requires a modification in the strategies for the continuous learning of the student [4].

3 New working environment

The University of Salamanca has a virtual environment, available for students and teachers, to incorporate new educative technologies to the development of educational tasks. The virtual campus, (`http://www.usal.es/eudored`), is based on a web platform called Moodle (Modular Object Oriented Distance Learning Environment), a course management system designed to help educators for creating quality online courses.

Moodle is a virtual environment for education which allows to place contents and tasks in the web and provides online communication tools. The design and development of Moodle is guided by a particular philosophy of learning: social constructionist Philosophy. With this learning philosophy people actively construct new knowledge as they interact with their environment, under the hypothesis that learning is more effective when you are constructing something.

One of the most important advantages of Moodle environment is that it has implemented all the useful tools and activities needed for online classes and e-Learning in general. The following features are part of the learning environment: The *Chat* module allows participants to have a real-time synchronous discussion via the web; in *forums* most discussion takes place, and they can be structured in different ways, and can include peer rating of each posting. Another activity are *glossaries*, that allows participants to create and maintain a list of definitions, like a dictionary. A module called *Hotpot* allows teachers to create multiple-choice, short-answer, jumbled-sentence, crossword, matching/ordering and gap-fill quizzes using Hot Potatoes software (`http://hotpot.uvic.ca/`). In Moodle platform *resources* can be prepared files uploaded to the course server; pages edited directly in Moodle; or external web pages made to appear part of this course. As part of web 2.0 learning tools, a *wiki* is a web site where anyone can add new contents or edit the existing ones, it enables documents to be authored collectively and supports collaborative learning.

4 Course activities: Training in Linear Algebra

We have used Moodle to create a new interactive educational teacher-student context. Students need to construct their own understanding of each algebraic concept, so that the primary role of teacher is not to explain, or attempt to 'transfer' knowledge, but to create situations for students that allow them to make the necessary mental constructions. In 21st century students are familiarized with the Internet and with the new technologies. They usually use them to chat with friends, to send and receive e-mails, to meet people or to organize holidays, but they are not conscious that it is a useful tool in the daily classes. Sometimes they do not see possible that personal computers and the Internet could be used effectively for classes about Mathematics.

With the purpose of obtaining a suitable training of the students, in each module we will give the students access to some interesting and introductory documentation, and we will create a forum to discuss about the current module. For example, with Systems of Linear Equations module we start a new Moodle activity which is a questionnaire with different items related to the right methods of solving systems of linear equations using matrices, or solving systems with some parameters. Other exercises will be proposed to the students so that they will be able to comment and debate them in the forums created for that goal. Moreover, some theoretical questions or Hot Potatoes exercises, that enable the creation of interactive tests, will be proposed for the students assessment.

Another interesting and practical exercise that we are planning is to propose the students to solve some problems as soon as possible. Each monday we will upload one problem and the first student who solve it will have an extra in their final assessment.

5 Conclusions

We have designed a new experience for teaching Linear Algebra in the University of Salamanca. The aim is to give the students a better understanding of that specific branch of Mathematics. In this paper we have proposed a web-based course according to the convergence of European Higher Education Project, to increase the use of new Information and Communication Technologies. This course will be available, for the students of the university, in the virtual environment, which is based on the Moodle platform, and offers a reachable environment easy to work with.

References

[1] T. Blackstone, *Education and Training in the Europe of Knowledge*, http://www.uniroma3.it/downloads/297_Lezione%20Blackstone.doc.

[2] Bologna Declaration http://www.ond.vlaanderen.be/hogeronderwijs/bologna/documents/MDC/BOLOGNA_DECLARATION1.pdf.

[3] A. García-Valcárcel Muñoz-Repiso, F. J. Tejedor Tejedor, *Current Developments in Technology-Assisted Education*, A. Méndez-Vilas, A. Solano Martín, J.A. Mesa González and J. Mesa González (eds.), FORMATEX, 2006.

[4] R. Mason, *Models of online courses*, ALN Magazine **2**, 2, 1998.

[5] A. Pérez i Garcías, *Nuevas estrategias didácticas en entornos digitales para la enseñanza superior*. En Didáctica y tecnología educativa para una univesidad en un mundo digital (J. Salinas y A. Batista), Universidad de Panamá, Imprenta universitaria, 2002.

[6] J. Salinas, *Innovación docente y uso de las TIC en la enseñanza universitaria*, Revista de Universidad y Sociedad del Conocimiento (RUSC) **1**, 1 (2004).

[7] J. Weiss, et al. (eds.) *The International Handbook of Virtual Learning Environments*, Springer, **14** (2006).

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Algorithms to encrypt and decrypt messages with Magma*

L. Hernández Encinas[1], J. Muñoz Masqué[1] and A. Queiruga Dios[2]

[1] *Dpt. Information Processing and Coding, Applied Physics Institute, CSIC*

[2] *Dpt. Applied Mathematics, University of Salamanca*

emails: luis@iec.csic.es, jaime@iec.csic.es, queirugadios@usal.es

Abstract

The security in current communications recommend to develop the implementation of cryptographic primitives and algorithms in an effective way. The cryptosystem proposed by Chor and Rivest, which is based on the knapsack problem, has recently been broken by Vaudenay but only when the original parameters are used. In this paper we give a brief overview of some developments in Cryptography and we present a safe implementation for Chor-Rivest Cryptosystem by using Magma software.

Key words: Chor-Rivest cryptosystem, Finite field, Implementation, Knapsack problem, Magma *software.*

1 Introduction

As it is well known, the goal of Cryptography is to guarantee the secrecy, confidentiality and integration of communications between several users. Moreover, the objective of Cryptanalysis is to break the security and privacy of such communications ([7], [8]). In public-key Cryptography each user has two different keys. One of them is the public key, which is publicly known and it is used by any sender to encrypt messages. The second key is the private key, which is kept in secret by the receiver and it is used by him to decrypt encrypted messages. In general, Public Key Cryptography bases its security on the computational intractability of some Number Theory problems, as factorization problem, discrete logarithm problem and knapsack problem.

An important public-key cryptosystem based on the knapsack problem was proposed by Chor and Rivest (see [3], [4]). This cryptosystem uses the arithmetic of finite

*Supported by Ministerio de Educación y Ciencia (Spain), under grant TSI2007-62657, and by CDTI, Ministerio de Industria, Turismo y Comercio (Spain) in collaboration with Telefónica I+D (Project SEGUR@) with reference CENIT-2007 2004.

fields and it needs to compute discrete logarithms over that field to determine the keys of the system. Nevertheless, the security of the system depends on the knapsack problem but not on the Discrete Logarithm Problem. In fact, if this problem becomes tractable, then the Chor-Rivest cryptosystem will be easier to implement.

The Discrete Logarithm Problem is nowadays considered a very difficult problem because the best algorithm known for solving it is the number field sieve ([10]) which has a subexponential expected running time:

$$O\left(e^{\left((64/9)^{1/3}+o(1)\right)(\ln p)^{1/3}(\ln\ln p)^{2/3}}\right).$$

The problem can be defined as follows: Given a prime integer p, a generator α of the group $\mathbb{Z}_p^*$, and an element $\beta\in\mathbb{Z}_p^*$, find an integer x, $0<x\leq p-1$, such that $\beta=\alpha^x$.

The Chor-Rivest cryptosystem has been broken by Vaudenay ([11]). Nevertheless, it was only broken for the original proposed parameters, i.e., over finite fields of q^h elements, $\mathbb{F}_{q^h}$, with $q\approx 200$ and $h\approx 25$, and h having a small divisor. In order to avoid Vaudenay's attack, it has been proved ([5]) that the values of the parameter h for which no divisor s verifies the equation $h/s+1\leq s$ are those for which h is either a prime number or the square of a prime number.

A secure Magma implementation of the Chor-Rivest cryptosystem is presented in this communication. This implementation uses new parameters, $q=409$, and $h=17$, which convert this cryptosystem in a safe system due to the fact that no attack has been presented when q and h are both prime numbers. The implementation includes procedures to transform the original messages in a suitable way, to generate the keys, and to encrypt and to decrypt messages in $\mathbb{F}_{q^h}$.

2 The Chor-Rivest cryptosystem implemented in Magma

We present the Magma procedures ([2]), functions, and statements needed to transform the messages, generate keys, encrypt, and decrypt messages using the Chor-Rivest cryptosystem, with a pair of real parameters safer than the original ones. Namely, we show the outputs of the implementation when the values $q=409$ and $h=17$, both prime numbers, are considered. That is, when the finite field is $\mathbb{F}_{409^{17}}$.

Magma ([1]) is a Computer Algebra System for solving problems in Algebra, Number Theory, Geometry and Combinatorics that may involve sophisticated mathematics and which are computationally hard. Magma provides a mathematically rigorous environment which emphasizes structural computation. Since the computation in Magma takes place in one or more algebraic structures, the first step in any computation involves defining those algebraic structures. Once the structures have defined, elements and other related objects of these structures may be created.

The parameters and the keys of the cryptosystem are chosen in the following way (see [3] and [4] for details):

1. Define the finite field $\mathbb{F}$ where the discrete logarithms must be calculated. The simplest way of creating the finite field (Galois field) with q elements in Magma

is to use the function `FiniteField` or `GF`, which are synonymous. To define the univariate transcendental extension $\mathbb{F}_q[t]$ of $\mathbb{F}_q$ the `PolynomialRing` command must be used.

2. Define the extension $\mathbb{F}_q$ of $\mathbb{F}$ by using the `ext` command, specifying a generator, g. In MAGMA the finite field $\mathbb{F}$ is built up as the extension of another finite field ($\mathbb{F}_q$ in our case), called the ground field, by an element called the generator of the field. The ground field is explicit in the construction when we use `ext`. An element of the field that generates the multiplicative group is called primitive. The function `IsPrimitive` may be used to check primitivity of an element. If $\mathbb{F}_q$ is a finite field and $f(t)$ is an irreducible polynomial, returned by the statement `IrreduciblePolynomial(Fq, h)` of degree h with coefficients in $\mathbb{F}_q$, this creates an extension of degree h of the ground field $\mathbb{F}_q$, with f as its defining polynomial.

3. Calculate the discrete logarithms. In this case, the best algorithm implemented in MAGMA to compute discrete logarithms for finite fields is Pohlig-Hellman algorithm ([9]), which running time is proportional to the square root of the largest prime l dividing n. This algorithm is always used for any finite field $\mathbb{F}$ if l is small and is also used if $\mathbb{F}$ is any non-prime field of characteristic greater than 2.

4. Calculate the noise as a random number in the interval $[0, q^h - 2]$ and determine a permutation of q elements.

5. With the previous procedures and statements, the user have determined the parameters of the system. Then, he computes and saves his private key (noise, r; permutation, π; and generator, g), and finally he computes the discrete logarithms a_i and determines the elements of his public key (the knapsack), c_i, which could be saved in a file and made public.

The public key of the user is the set $(c_0, c_1, \ldots, c_{q-1})$, and his/her private key is the set (t, g, π, r).

Moreover, for this system, the messages must be binary vectors of length q and weight h. So, we suppose that this is the form of a message.

In addition to the previous commands, other commands must to be used. For example, `Eltseq` command, which converts the input parameter of the finite field $\mathbb{F}$ and its representation as a polynomial in the generator of $\mathbb{F}$ over a subfield $\mathbb{F}_q$, with coefficients in $\mathbb{F}_q$; `CodeToString` and `StringToCode` commands for converting between a one-character string and the character code that it has in the computers operating system; and `Intseq(a, b)` command which gives the sequence $[a_0, a_1, \ldots, a_n]$ which is the representation of $a = a_0b^0 + a_1b^1 + \ldots + a_nb^n$ in the base b.

In [6] an implementation by using MAPLE software was presented, but it is less efficient than the one shown here. In both implementations, most of the time of CPU is spent in the calculation of discrete logarithms. In the example presented in [6], the process of generation of the parameters and keys required 76.42 minutes, but with our implementation in MAGMA this time was reduced to 2.22 minutes on a Intel Core 2 CPU T7200 2.00 GHz, 2.00 GB RAM.

3 Conclusions

We have studied and implemented the Chor-Rivest cryptosystem with Magma software. The implementation of this cryptosystem was made with safe parameters, that is, $q = 409$, h017 in an efficient way. This implementation includes functions to transform the messages, procedures to generate the keys, encrypt and decrypt messages in the finite field $\mathbb{F}_{q^h}$. The implementation opens a door to the future in order to use the Chor-Rivest cryptosystem in a realistic way due to the fact that no attack has yet been proposed when the parameters q and h are both prime numbers.

References

[1] W. Bosma, J. Cannon, and C. Playoust. *The Magma algebra system. I. The user language*. J. Symbolic Comput., 24, 3-4 (1997), 235–265.

[2] J. J. Cannon, W. Bosma (Eds.) *Handbook of Magma Functions*, Edition 2.12, School of Mathematics and Statistics, University of Sydney, 2006.

[3] B. Chor, *Two issues in public key cryptography. RSA bit security and a new knapsack type system*, The MIT Press, Cambridge, MS, 1985.

[4] B. Chor and R.L. Rivest, *A knapsack-type public key cryptosystem based on aritmethic in finite fields*, IEEE Trans. Inform. Theory **34**, 5 (1988), 901–909.

[5] L. Hernández Encinas, J. Muñoz Masqué, and A. Queiruga Dios, *Análisis del criptosistema de Chor-Rivest con parámetros primos*, Actas de la IX Reunión Española sobre Criptología y Seguridad de la Información (IX RECSI), 548–561, Barcelona (2006).

[6] ———, *Maple implementation of the Chor-Rivest cryptosystem*, Lecture Notes in Comput. Sci. **3992** (2006), 438–445.

[7] A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of applied cryptography*, CRC Press, Boca Raton, FL, 1997.

[8] R.A. Mollin, *An introduction to cryptography*, Chapman & Hall/CRC, Boca Raton, FL, 2001.

[9] R.C. Pohlig and M.E. Hellman, *An improved algorithm for computing logarithms over $GF(p)$ and its cryptographic significance*, IEEE Trans. Inform. Theory **24** (1978), 106–110.

[10] O. Schirokauer, D. Weber, and T. Denny, *Discrete logarithms: the effectiveness of the index calculus method*, Algorithmic Number Theory, LNCS **1122** (1996), 337–361, Springer-Verlag, Berlin.

[11] S. Vaudenay, *Cryptanalysis of the Chor-Rivest cryptosystem*, J. Cryptology **14** (2001), 87-100.

Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

CsegGraph: A Graph Coloring Instance Generator in Sparse Derivative Matrix Determination

Shahadat Hossain[1]

[1] *Department of Mathematics and Computer Science, University of Lethbridge*
emails: `shahadat.hossain@uleth.ca`

Abstract

A graph generator associated with the determination of mathematical derivatives is described. The graph coloring instances are obtained as intersection graphs $G_\Pi(A)$ of the sparsity pattern of $A \in \Re^{m\times n}$ with row partition Π. The size of the graph is dependent on the row partition; the number of vertices can be varied between the number of columns (using single block row partition $\Pi = \Pi_1$) and the number of nonzero entries of A (using m block row partition $\Pi = \Pi_m$). The chromatic number of the generated graph instances satisfy $\chi(G(A)) \equiv \chi(G_{\Pi_1}(A)) \leq \chi(G_\Pi(A)) \leq \chi(G_{\Pi_m}(A))$.

Key words: Column Segment Graph, Sparse Matrix, Mathematical Derivative.

1 Introduction

Graph coloring problems arise in a variety of scientific applications and are one of the widely studied class of problems in graph theory. Applications where the underlying problem is modelled by the coloring of the vertices of a graph arise, for example, in scheduling and partitioning problems, matrix determination problems [2, 5], and register allocation problems. Unfortunately, determining whether or not an arbitrary graph is p-colorable $p \geq 3$ is NP-complete [7]. The availability of suitable benchmark test problems is therefore an important component in the design and testing of effective algorithms for graph coloring and related problems. The main purpose of this paper is to describe an implementation of graph coloring test instances described in [6]. Some of the hardest problem instances included in the DIMACS graph coloring benchmark instances (`http://mat.gsia.cmu.edu/COLORING02/` accessed May, 2008) are obtained from the graph generator presented here.

The remainder of this paper is organized in five sections. Section 2 provides a brief introduction to the coloring problem associated with sparse derivative matrix determination. In section 3, an algorithmic description of the column-segment graph generator is given. In section 4 we present a graph generator based on a partition of the edges of an undirected graph. The instructions for using our graph generator is given in the Appendix. Section 5 concludes the paper.

2 The Coloring Problem

A *Graph* $G = (V, E)$ is a finite set V of *vertices* and a set E of *edges*. An edge $e \in E$ is denoted by an unordered pair $\{u, v\}$ which connects vertices u and v, $u, v \in V$. A graph G is said to be a *complete* graph or a *clique* if there is an edge between every pair of distinct vertices. In this paper multiple edges between a pair of vertices are considered as a single edge. A *p-coloring* of the vertices of G is a function $\Phi : V \mapsto \{1, 2, \cdots, p\}$ such that $\{u, v\} \in E$ implies $\Phi(u) \neq \Phi(v)$. *The chromatic number* $\chi(G)$ of G is the smallest p for which it has a *p-coloring*. An *optimal coloring* is a *p-coloring* with $p = \chi(G)$.

Given $A \in \Re^{m\times n}$, the *intersection graph* of the columns of A is denoted by $G(A) = (V, E)$ where corresponding to column j of A, written $A(:, j), j = 1, 2, \ldots, n$, there is a vertex $v_j \in V$ and $\{v_j, v_l\} \in E$ if and only if there is a row index $1 \leq i \leq m$ for which $a_{ij} \neq 0$ and $a_{il} \neq 0$, $l \neq j$.

A *row $\tilde{q}$-partition* Π is a partition of $\{1, 2, \ldots, m\}$ yielding $w_1, w_2, \ldots, w_{\tilde{i}}, \ldots, w_{\tilde{q}}$ where $w_{\tilde{i}}$ contains the row indices that constitute the *block* $\tilde{i}$, denoted by $A(w_{\tilde{i}}, :) \in R^{m_{\tilde{i}}\times n}$, $\tilde{i} = 1, 2, \ldots, \tilde{q}$. A segment of column j in block $\tilde{i}$ of A denoted by $A(w_{\tilde{i}}, j), \tilde{i} = 1, 2, \ldots, \tilde{q}$ is called a *column segment.* Unless explicitly stated the column segments in the following are not identically zero.

Definition 2.1 (Structurally Orthogonal Column Segments).

- (`Same Column`)

  Column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{k}}, j)$, $\tilde{i} \neq \tilde{k}$ are *structurally orthogonal.*

- (`Same Row Block`)

  Column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{i}}, l)$, $j \neq l$ are *structurally orthogonal* if they do not have nonzero entries in the same row position.

- (`Different`)

  Column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{k}}, l)$, $\tilde{i} \neq \tilde{k}$ and $j \neq l$ are *structurally orthogonal* if

  - $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{i}}, l)$ are *structurally orthogonal* and
  - $A(w_{\tilde{k}}, j)$ and $A(w_{\tilde{k}}, l)$ are *structurally orthogonal.*

An *orthogonal partition of column segments* is a mapping

$$\kappa : \{(\tilde{i}, j) : 1 \leq \tilde{i} \leq \tilde{q}, 1 \leq j \leq n\} \mapsto \{1, \ldots, p\}$$

such that column segments in each group in the partition are structurally orthogonal.

Definition 2.2 (Column-segment Graph). Given matrix A and row $\tilde{q}$-partition Π, the *column-segment graph* associated with A under Π is a graph $G_\Pi(A) = (V, E)$ where vertex $v_{\tilde{i}j} \in V$ corresponds to the column segment $A(w_{\tilde{i}}, j)$ not identically 0, and $\{v_{\tilde{i}j}, v_{\tilde{k}l}\} \in E$ $1 \leq \tilde{i}, \tilde{k} \leq \tilde{q}, 1 \leq j, l \leq n$ if and only if column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{k}}, l)$ are not structurally orthogonal.

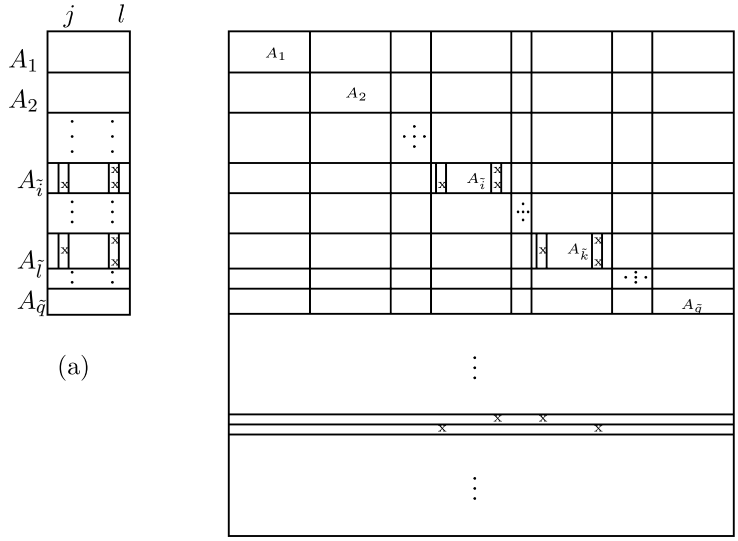
The problem of determining a Jacobian matrix A using column segments can be stated as the following graph problem.

Theorem 2.3. *[5] Φ is a coloring of $G_\Pi(A)$ if and only if Φ induces a orthogonal partition κ of the column segments of A.*

3 The Column Segment Graph Generator

In this section we describe an algorithm for constructing *column segment graph* $G_\Pi(A)$ associated with a $m \times n$ matrix A and a row partition Π. Furthermore, we describe the column segment matrix A_Π associated with the given row partition.

Let Π be a row partition of matrix A that partitions the rows into blocks $A_1, A_2, \ldots, A_{\tilde{q}}$ (See Fig. 1(a)). Denote the intersection graph corresponding to $A_{\tilde{i}}$ by $G(A_{\tilde{i}}), \tilde{i} = 1, 2, \ldots \tilde{q}$. The construction of A_Π involves two phases. In the first phase, blocks $A_{\tilde{i}}$, $\tilde{i} = 1, 2, \ldots, \tilde{q}$ are placed successively in the left-to-right fashion (see Fig. 1(b)) such that each nonzero column segment is mapped to a unique column of A_Π. In other words, for every nonzero column segment of A, a column is created in A_Π where all the entries are zero except that the column segment is copied in the matching row positions. This situation is illustrated in the top part of Fig. 1(b). In the second phase of construction, restrictions are enforced on column segments that are not structurally orthogonal in order to prevent them from being grouped together. To illustrate, consider the column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{i}}, l)$ in $A_{\tilde{i}}$. If there are nonzero entries in the same row posi-



(a)

(b)

Figure 1: (a): Matrix A is partitioned into $\tilde{q}$ blocks (b): Column segment matrix corresponding to the partition.

tion in $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{i}}, l)$ then they are not orthogonal implying that $A(w_{\tilde{i}}, j)$ is not orthogonal to column segments $A(w_{\tilde{k}}, l)$ for all $\tilde{k} \neq \tilde{i}$. Consequently, $A(w_{\tilde{i}}, j)$ cannot be grouped together with any of the segments $A(w_{\tilde{k}}, l)$. Similarly, $A(w_{\tilde{i}}, l)$ is not orthogonal to columns $A(w_{\tilde{k}}, j)$ for all $\tilde{k} \neq \tilde{i}$. To enforce these restrictions we simply introduce two new rows in A_Π, one containing nonzero entries in the column positions mapped by the column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{k}}, l)$ and the other containing nonzero entries in the column positions mapped by the column segments $A(w_{\tilde{i}}, l)$ and $A(w_{\tilde{k}}, j)$ for all $\tilde{k} \neq \tilde{i}$. This is done for every pair of dependent column segments in $A_{\tilde{i}}$, $\tilde{i} = 1, 2, \ldots, \tilde{q}$ (see Fig. 1(b)). To see this dependency restriction in terms of graphs, consider vertices $v_{\tilde{i}j}$ and $v_{\tilde{i}l}$ in $G(A_{\tilde{i}})$. For each such edge we define edges between vertex $v_{\tilde{i}j}$ and vertices $v_{\tilde{k}l}$ from $G(A_{\tilde{k}})$ for $\tilde{k} \neq \tilde{i}$. Similarly, vertex $v_{\tilde{i}l}$ is connected with the vertices $v_{\tilde{k}j}$ from $G(A_{\tilde{k}})$ for $\tilde{k} \neq \tilde{i}$. This situation is illustrated in Fig. 2. In Fig. 1(a) the matrix is partitioned into $\tilde{q}$ blocks denoted by $A_1, A_2, \ldots, A_{\tilde{q}}$. In Fig. 1(b) placement of each of the blocks $A_1, A_2, \ldots, A_{\tilde{q}}$ is shown. Column segments $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{i}}, l)$ are not orthogonal, and hence two rows containing nonzero entries in the appropriate columns of $A_\Pi$are introduced.

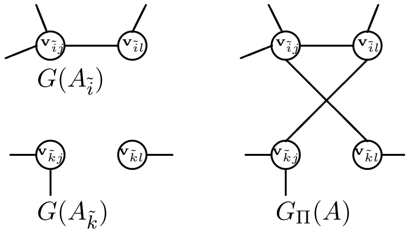


Figure 2: Graph $G_\Pi(A)$ before and after the insertion of edges due to the edge $\{v_{\tilde{i}j}, v_{\tilde{i}l}\}$ in $G(A_i)$.

From the procedure for column segment matrix construction described above we obtain the following result.

Theorem 3.1. *[5] $G(A_\Pi)$ is isomorphic to $G_\Pi(A)$.*

An upper bound on the size of the column segment matrix and graph is easily obtained from its construction. For a $\tilde{q}$ partition, the number of columns $n' \leq nnz \leq n * \tilde{q}$ where nnz is the number of nonzero elements in A. The number of rows $m' \leq m + (\tilde{q} - 1) \sum_{i=1}^{m} \rho_i(\rho_i - 1)$ where ρ_i is the number of nonzero in the ith row of A. In practice, however, the numbers n' and m' are smaller due to many zero column segments and repeated edges between pair of distinct vertices.

4 A Graph Theoretic Approach

The graph generator of the preceding section is based on row partition of sparse pattern matrices. The generated graph instances are described by listing the edges (undirected) followed by the number of vertices and edges of the graph. An instance generator can also be described in purely graph-theoretic terms. Let $G = (V, E)$ be an undirected graph with $|V| = n > 0$ vertices and $|E| = m > 0$ edges. Let Π be a partition[1] of the edges of E into subsets $E_1, E_2, \ldots, E_{\tilde{q}}$. Define graphs $G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_{\tilde{q}} = (V_{\tilde{q}}, E_{\tilde{q}})$ where $V_{\tilde{i}} = \{v \in V|$ there is an edge $e \in E_{\tilde{i}}$ which is incident on $v\}, \tilde{i} = 1, 2, \ldots, \tilde{q}$. A construction similar to the one shown in Figure 2 can be used to introduce new edges to incorporate dependency information among the subgraphs $G_{\tilde{i}}, \tilde{i} = 1, 2, \ldots, \tilde{q}$.

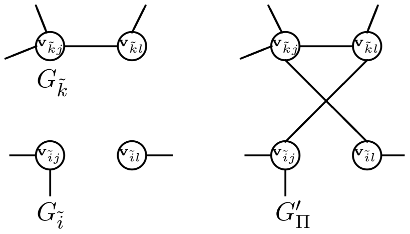


Figure 3: Graph G'_{Π} before and after the insertion of edges due to the edge $\{v_{\tilde{k}j}, v_{\tilde{k}l}\}$ in $G_{\tilde{k}}$.

In Figure 3 the subgraph $G_{\tilde{k}}$ contains edge $\{v_{\tilde{k}j}, v_{\tilde{k}l}\}$. This dependency is incorporated in the graph G'_{Π} by defining the "dependency edges" $\{v_{\tilde{k}j}, v_{\tilde{i}l}\}$ and $\{v_{\tilde{k}l}, v_{\tilde{i}j}\}$ for each subgraph $G_{\tilde{i}}, \tilde{i} = 1, 2, \ldots, \tilde{q}$. Let $\widehat{E}$ be the set of such dependency edges. Then the resulting graph

$$G'_{\Pi} = (V', E')$$

where

$$V' = \bigcup_{\tilde{i}} V_{\tilde{i}}, \text{ and } E' = \left(\bigcup_{\tilde{i}} E_{\tilde{i}}\right) \bigcup \widehat{E}, \tilde{i} = 1, 2, \ldots, \tilde{q}$$

satisfies $\chi(G'_{\Pi}) \leq \chi(G)$.

Denote by $G'_{\Pi}(A) = (V', E')$ the "extended graph" of $G(A)$ (with $\tilde{q}$ block row partition Π) which is obtained via a simple modification of definition 2.2 where corresponding to each column segment, including the identical zero column segment, there is a vertex in V'. Let $G_{\Pi}(A) = (V, E)$ be the column segment graph under the same

[1] $$E = \bigcup_{\tilde{i}} E_{\tilde{i}}, \;\; E_{\tilde{i}} \cap E_{\tilde{j}} = \emptyset \text{ whenever } \tilde{i} \neq \tilde{j}, \text{ and } E_{\tilde{i}} \neq \emptyset, \tilde{i} = 1, 2, \ldots, \tilde{q}.$$

row partition Π. Then we have the following result that $G_\Pi(A)$ is p-colorable if and only $G'_\Pi(A)$ is p-colorable. To see this let Φ be a p-coloring of $G_\Pi(A)$. We show that Φ can be modified to construct a new p-coloring Φ' for $G'_\Pi(A)$. First, let $\Phi'(v_{\tilde{i}j}) = \Phi(v_{\tilde{i}j})$ corresponding to nonzero column segments $A(w_{\tilde{i}}, j), \tilde{i} = 1, 2, \ldots, \tilde{q}$ and $j = 1, 2, \ldots, n$. Let $A(w_{\tilde{k}}, q)$ be any column segment which is identical zero. Then $\{v_{\tilde{i}j}, v_{\tilde{k}l}\} \in E'$ implies that $\tilde{i} \neq \tilde{k}$ and $j \neq l$. Consequently, $A(w_{\tilde{i}}, j)$ and $A(w_{\tilde{i}}, l)$ must be nonzero and the corresponding vertices are included in V. We set $\Phi'(v_{\tilde{k}l}) = \Phi(v_{\tilde{i}l})$. Since $\Phi(v_{\tilde{i}j}) \neq \Phi(v_{\tilde{i}l})$ the coloring of $v_{\tilde{k}l}$ is valid. Since $G_\Pi(A) \subset G'_\Pi(A)$, Φ is a p-coloring of $G'_\Pi(A)$ implying that Φ is a p-coloring of $G_\Pi(A)$. Therefore, $G_\Pi(A)$ is p-colorable if and only $G'_\Pi(A)$ is p-colorable.

We obserbe that given a connected undirected graph $G = (V, E)$ we can construct $A \in \{0, 1\}^{|E| \times |V|}$ such that $G(A)$ is isomorphic to G. The preceding discussion on the coloring of the graphs $G_\Pi(A)$ and $G'_\Pi(A)$ together with the above observation outlines a framework for the interpretation of the graph generator from a purely graph-theoretic view point.

5 Concluding Remarks

In this paper we have described a graph instance generator based on intersection graphs of row partitioned sparse pattern matrices. We have outlined a procedure for defining graph instances based on edge partition of a given input graph. That the generated instances are highly structured and the size of the generated instances can be varied easily make them convenient for use as test sets for combinatorial optimization problems such as graph coloring.

Acknowledgements

The author is grateful to Professor Trond Steihaug who read an early draft of the manuscript and provided suggestions for improvement.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under a Discovery Grant.

6 Appendix A (Graph Generator Usage)

Our graph generator implements column segment graph instances. The software uses SparseLib++v.1.5d [3], a collection of C++ sparse matrix classes that can read and convert between a number of standard sparse matrix data structures e.g., coordinate, compressed column, and compressed row format which are also supported by Harwell-Boeing test matrix collection [4].

The C++ source code is provided in two directories:

1. The directory named `col_seg_graph` contains C++ code implementing the column-segment matrix and the associated graph from a given sparse matrix. The di-

rectory also contains several utility functions and a Makefile that can be used to generate the executables.

2. The directory named `SparseLib++` contains the sparse matrix library described in [3]. `SparseLib++` provides support for Harwell-Boeing and Matrix Market [1] sparse matrix exchange formats.

6.1 Column Segment Graph

The function for defining column segment graph has the following prototype declaration.

```
Coord_Mat_double& extend( const Coord_Mat_double& A,
                          const vector<int>& perm,
                          const vector<int>& part,
                          const char* fileName );
```

Given input matrix A in `Coord_Mat_double` format [1], input vector `perm` representing a permutation of the row indices of A, input vector `part` representing a row partition of A, and character string `fileName`, `extend` returns the resulting column segment matrix in `Coord_Mat_double` format as function return value, and writes the associated column segment graph in the file `fileName`.

Users can make (by running the command `make`) the executable program named `extend` which can be executed to generate column segment graph and the associated matrix.

```
Usage: extend TESTFILE [OUT_FILE (Optional)]
```

Information such as row partition, permutation, and the location of the input matrix are provided in an input text file `TESTFILE`. Argument `OUT_FILE_NAME` stores the column segment graph and is optional; if not provided the graph is written to the standard output. The resulting column segment matrix is stored in a text file called `inputMatrixFileName_ext.mtx` where `inputMatrixFileName.mtx` is the name of the input matrix in Matrix Market exchange format.

The format of `TESTFILE` is described below.

Comment lines: The hash sign (#) at the beginning of a line marks that line as comment. The comments can only appear at the beginning of the file (i.e., before permutation and partition data) and cannot be interleaved with data.

The partition size is a single integer that specifies the number of row blocks in the partition.

Partition lines: Following the comment lines and partition size, commences the specification of the row partition given by listing the index of the first row of each row block: $r_1\ r_2\ \ldots\ r_{nblks+1}$ where $nblks$ denotes the number of row blocks in the partition. The first row block in the partition consists of rows $r_1, \ldots, r_2 - 1$, the second row block in the partition consists of rows $r_2, \ldots, r_3 - 1$ and so on. Since a permutation of the rows can also be specified (explained next) the indices r_1, r_2 etc. are given in increasing order with $r_1 = 1$ and $r_{nblks} + 1 = m + 1$ where m denotes the number of rows in the input matrix. An m block partition can also be indicated by writing the negative of the integer m+1.

Permutation Lines: It is possible to specify a permutation of rows in specifying the row partition. This allows for the rows in a row block not necessarily be consecutive. For example, if we have 4 rows and the row partition is 1 3 5 and the permutation is given as 3 1 2 4 then first block consists of rows with indices 3 1 and second block consists of rows with indices 2 4. If no permutation need to be specified then the negative of the number of rows m is written in the permutations lines.

Input Matrix file: This last line specifies the name of the file (full path name) containing the input matrix.

6.2 Examples

Example test file (`TESTFILE`) Example 1:

```
# File t1.input
# Input matrix has 4 columns and 4 rows
# Row 2-partition with permutation
    2
    1 3 5
    2 3 1 4
    input_dir/test1.mtx
```

The first 3 lines are comments. The fourth line says that the row partition defines two blocks. The fifth line specifies the two-block row partition: the first block starting at row 1 and ending at row 2, the second starting at row 3 and ending at row 4. There are 4 rows in the input matrix. The sixth line says that a row permutation is provided: rows 2 and 3 of the input matrix constitute the first block and the rows 1 and 4 constitute the second block. The seventh line specifies that the input matrix is contained in file `test1.mtx` in the directory `input_dir`. The suffix `mtx` indicates that the input sparse matrix is provided in Matrix Market format. The input matrix can also be provided in Harwell-Boeing format (`test1.p[rsu][ae]`). Note that a Harwell-Boeing pattern matrix has a three letter suffix in which the first one is the character `p`, the second is one of the characters `r,s`, or `u`, and the third character is either an `a` or an `e`. (The current implementation only supports matrix market exchange format for output of column segment matrix.) The content of the input matrix `test1.mtx` is shown below.

```
%%MatrixMarket matrix coordinate pattern general
4 4 8
1 1
```

```
1 2
2 1
2 3
3 2
3 3
4 3
4 4
```

Then the command

```
extend test1.input test1_ext.graph
```

will create the files `test1_ext.mtx` which contains the column segment matrix and `test1_ext.graph` contains the column segment graph corresponding to the given row partition.

Content of `test1_ext.mtx`:

```
%%MatrixMarket matrix coordinate pattern general
% Generated by writeMM()
11 7 22
  1     1
  1     3
  2     2
  2     3
  3     4
  3     5
  4     6
  4     7
  5     1
  5     6
  6     3
  6     4
  7     2
  7     6
  8     3
  8     5
  9     4
  9     2
 10     5
 10     1
 11     7
 11     3
```

Content of `test1_ext.graph`

```
  1     3
```

```
   2      3
   4      5
   6      7
   1      6
   3      4
   2      6
   3      5
   4      2
   5      1
   7      3

 7 11
```

Figure 4 displays the column intersection graph and the column segment graph of the example contained in `test1.mtx` (A). An edge of the original graph $G(A)$ is indicated by a solid line. The dashed lines in the column segment graph denote edges introduced to enforce dependency. Note that in Figure 4 $G_{\Pi}(A)$ has seven vertices and the indices of these vertices are indicated in parentheses. In the graph file `test1_ext.graph` the edges of the column segment graph are given as pairs of indices one edge per line in all but the last line. The last line contains the number of vertices (7) and the number of edges (11) of the column segment graph. The column segment matrix contained in file `test1_ext.graph` is described using matrix market format.

Example test files (`TESTFILE`) Example 2:

```
# File t2.input
# Input matrix has 4 columns and 4 rows
# Row 4-partition and no permutation
    4
    -5
    -4
    input_dir/test2.mtx
```

The first 3 lines are comments. The fourth line says that this row partition has four blocks: each row constitutes a block. The input matrix has the same dimension as in Example 1. The fifth line says that no row permutation is given which is indicated by the negative of the $m + 1$ where $m = 4$ denotes the number of rows of the input matrix. The sixth line says that no row permutation is provided. The last line specifies that the input matrix is contained in file `test2.mtx` in the directory `input_dir`.

6.3 Utilities

The graph generator is packaged with a number of utility programs for user convenience. The executable `showmat` displays small (with less than approximately 30 columns) pattern matrices on to the terminal screen.

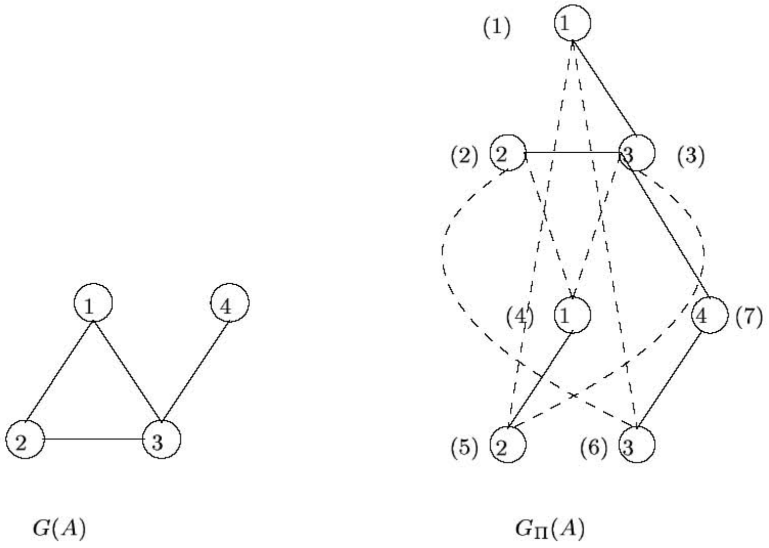


Figure 4: The column intersection graph $G(A)$ and the column segment graph $G_\Pi(A)$ for the test matrix A provided in test1.mtx

```
Usage: showmat { -m MFILE | -t TESTFILE }
            -m: MFILE is a matrix in Matrix Market or
            Harwell-Boeing exchange format
            -t: TESTFILE is an input file as in extend
```

With option `-m` the argument is expected to be a file that describes a sparse pattern matrix in either Matrix Market or Harwell-Boeing format. Option -t, on the other hand, allows users to specify a `TESTFILE` (see the usage of `extend` command).

Since the column segment graph output by `extend` does not conform to the input format of any particular graph coloring application we provide Perl scripts to format the column segment graph. The script `ToDSaturFmt.pl` converts the column segment graph to the input format of `DSATUR` graph coloring implementation by Michael Trick [8].

```
 Usage:Perl ToDSaturFmt.pl InputFile OutPutFile
       InputFile is a ASCII file describing
       column segment graph (output of extend function)
       OutPutFile is a ASCII file describing Column Segment
       graph in the input format for DSATUR implementation by
       Michael Trick.
```

The script `ToDIMACS.pl` converts the column segment graph to the input format of DIMACS challenge.

```
Usage:Perl ToDIMACS.pl InputFile OutPutFile
      InputFile is a ASCII file describing
      column segment graph (output of extend function)
      OutPutFile is a ASCII file describing Column Segment
      graph in the input format for DIMACS challenge.
```

References

[1] R. F. Boisvert, R. Pozo, K. Remington, R. Barrett, and J. J. Dongarra. The Matrix Market: A web resource for test matrix collections. In R. F. Boisvert, editor, *Quality of Numerical Software, Assessment and Enhancement*, pages 125–137, London, 1997. Chapman and Hall.

[2] T. F. Coleman and J. J. Moré. Estimation of sparse Jacobian matrices and graph coloring problems. *SIAM J. Numer. Anal.*, 20(1):187–209, 1983.

[3] J. Dongarra, A. Lumsdaine, R. Pozo, and K. Remington. A sparse matrix library in c++ for high performance architectures. In *Proceedings of the Second Object Oriented Numerics Conference*, pages 214–218, 1994.

[4] I. S. Duff, R. G. Grimes, and J. G. Lewis. Sparse matrix test problems. *ACM Transactions on Mathematical Software*, 15(1):1–14, 1989.

[5] S. Hossain and T. Steihaug. Optimal Direct Determination of Sparse Jacobian Matrices. Technical Report 254, Department of Informatics, University of Bergen, Norway, October 2003 (Revised version to appear in Optimization Methods and Software).

[6] S. Hossain and T. Steihaug. Graph coloring in the estimation of sparse derivative matrices: Instances and applications. *Discrete Appl. Math. 156(2):280–288, 2008.*

[7] M.R.Garey and D.S.Johnson. *Computers and Intractibility.* W.H.Freeman, San Francisco, 1979.

[8] A. Mehrotra and M. A. Trick. A column generation approach for graph coloring. *INFORMS Journal on Computing*, 8:344–354, 1996.