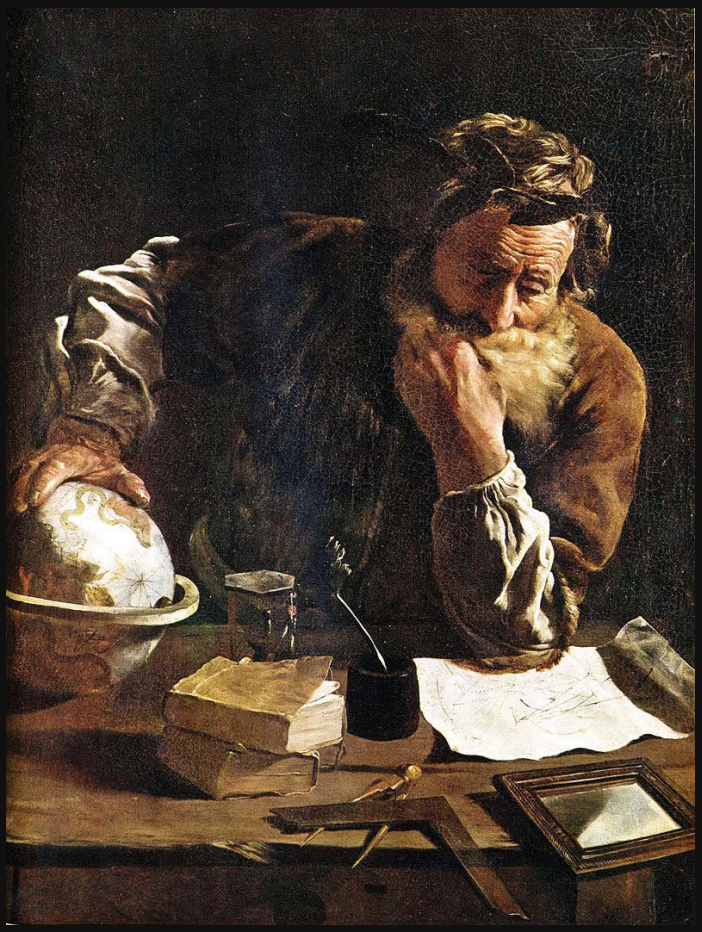
Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering

Costa Ballena (Rota), Cádiz, Spain
July 4th-8th, 2016



Editors

J. Vigo-Aguiar

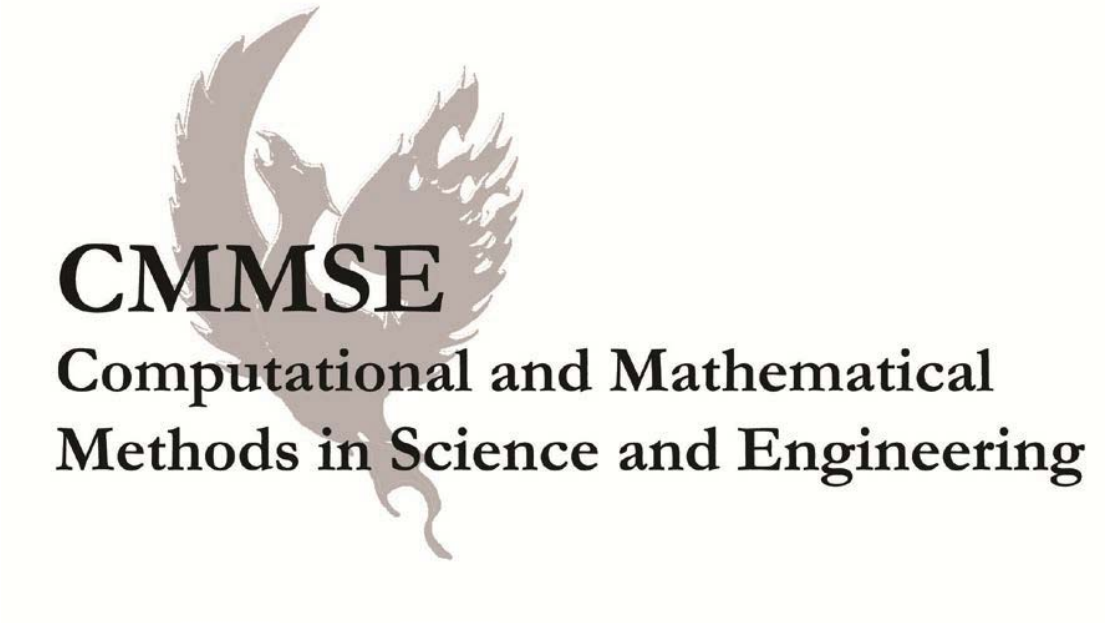
Associate Editors

P. Schwerdtfeger (New Zealand), W. Spröβig (Germany), N. Stollenwerk (Portugal), Pino Caballero (Spain), J. Cioslowski (Poland), J. Medina (Spain), I. P. Hamilton (Canada), J.A. Alvarez-Bermejo (Spain)

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering CMMSE-2016

Costa Ballena (Rota), Cádiz, Spain

July 4-8, 2016



Editors

J. Vigo-Aguiar

Associate Editors

P. Schwerdtfeger, W. Sprößig, N. Stollenwerk, Pino Caballero, J. Cioslowski, J. Medina, I. P. Hamilton, J.A. Alvarez-Bermejo

ISBN **978-84-608-6082-2**

Printed on acid-free paper

Cover: Arquímedes by Domenico Fetti

These volumes are dedicated to the families of some CMMSE members, by their patience and understanding all these years (Raquel García, Claudia & Víctor Vigo, Ana Fernández, Fernando Alonso, Sandra & Paola Ranilla).

Contents:

Volume I:

Volume II:

Volume III:

Volume IV:

Volume V:

Late Contributions:

Preface

It is a great pleasure to welcome you to the **16th International Conference on Computational and Mathematical Methods in Science and Engineering** (CMMSE 2016), will be held at Rota, Cádiz (Spain), July 4th-8th, 2016. This consolidate international conference offers the opportunity of discuss on new breakthrough ideas in different applied mathematics disciplines and creates synergies among the attendants in order to advance the frontier of knowledge and, as a consequence, to propose new (European) research projects.

Horizon 2020 is the new EU Framework Programme for Research and Innovation. New Societal and Technology problems have been raised where Applied Mathematics seems to be a fundamental field in the vast majority of them. The importance of this discipline in other areas, such as Engineering, Computer Science, Physics and Chemistry, must be a value in order to be part of the funded projects.

The different advances are included in the extended abstracts and papers accepted to the conference and will be collected in these proceedings of CMMSE 2016. The proceedings we have pleasure to present here has five volumes, the first four correspond to the articles typeset in LaTeX and the fifth to articles typeset in Word. During CMMSE 2016 you will have the opportunity of discuss about current and new challenges, exchanging ideas, comments and suggestions leading to the improvement and deepening of the papers presented to allow further development of the research occurs. The attendants must also take advantage of the plenary speakers and important attendant researchers, in order to get new ideas and, maybe, propose future collaborations.

Twenty symposia show the variety of disciplines considered in the conference, which are formed from high quality accepted papers. The first one, *high-performance computing*, considers new large-scale problems that arise in fields like bioinformatics, computational chemistry, and astrophysics. *Mathematically modeling the future Internet and developing future Internet security technology* is a self-explanatory session. The third symposium addresses analytical, numerical and computational aspects of *partial differential equations in life and materials science*. *Computational finance* is a session focusing on solving problems related to asset pricing, trading and risk analysis of financial assets that have no analytic solutions under realistic assumptions and thus require computational methods to be resolved. A forum for discussion of the growing impact of new technologies on teaching and the development of new tools to increase learning efficiency is provided in the symposium: *new educational methodologies supported by new technologies*. The symposium on *mathematical models and information-intelligent systems on transport* researches in the field of flow-modelling of particles with motivated behavior in complex networks, applied to traffic flows, pedestrian flows, ecology, etc. The seventh symposium studies *computational methods for linear and nonlinear optimization* and *numerical methods for solving nonlinear problems* is given in another session. *Bio-mathematics* studies both theoretical and practical applications of population dynamics, eco-epidemiology, epidemiology of infectious diseases and molecular and antigenic evolution. The 10th symposium will put particular emphasis on species involving Coulombically interacting particles either trapped by external potentials or confined to hypersurfaces: *Quasi-Solvable Systems in Quantum Chemistry and Physics*. Model interesting problems arisen in Computer Science, considering algebraic and computational (fuzzy) techniques, is the main goal of *mathematical models for computer science*. The aim of the 12th

symposium is to obtain a consistent description of the transition *from clusters to the solid state*, which is a major challenge in computational chemistry and physics.

The enormous potential of fixed point theory, which is needed in mathematics, engineering, chemistry, biology, economics, computer science, and other sciences, justifies the great interest in *fixed point theory in various abstract spaces and related applications*. *Computational methods in direct and inverse (systems of) PDE's* covers general phenomena formulated as control problems or inverse problems associated with mathematical models described by partial differential equations (PDE). An overview of mathematical and computational research focusing on corporate or government applications and problems arising from different economic sectors is presented in the 16th mini-symposium: *Industrial Mathematics*. Parallel implementation using hybrid architectures with accelerators, either GPUs or FPGAs, of numerical methods for solving problems within the following topics of interest: industrial mathematics, fluid mechanics, global optimization, finance, geophysical flows, computational chemistry, electromagnetism, magneto hydrodynamics, atomic physics, relativistic flows; is given in the symposium: *Numerical simulation on GPUs*. *Mathematics in the Information Society* aims at presenting recent advances in practical applications of Cryptology, Coding Theory and Development of Technologies that address challenges related to Communication Systems. The *computational methods for fluid flow* are considered in this symposium, which provides a forum for discussion of current problems and recent advances in the area. Finally, the last symposium, *Uncertainty and Imprecision Modelling in Decision Making,* introduces new results about the concepts and procedures commonly used in decision-making processes where an imprecise knowledge of the data is present.

We want to conclude this preface giving thanks to the plenary speakers for their outstanding contributions to research and leadership in their respective fields, including physics, computer science and engineering. We would also like to thank the special session organizers and scientific committee members, who have played a very important part in setting the direction of CMMSE 2016. Finally, we would like to thank the participants because, without their interest and enthusiasm, the conference would not have been possible.

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

Costa Ballena, Rota, Cádiz (Spain), July 3rd, 2016

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

CMMSE 2016 Mini-symposia

Session Title	Type of Session/Organizers
High Performance Computing (HPC)	J. Ranilla
P.D.E.'S in Life and Material Sciences	Regular Session
Computational Finance	
New Educational Methodologies Supported by New Technologies	Regular Session
Mathematical Models and Information-Intelligent Systems on Transport	Alexander P. Buslaev & M. Yashenina
Computational Methods for Linear and Nonlinear Optimization	Maria Teresa Torres Monteiro
Numerical Methods for Solving Nonlinear Problems	Juan R. Torregrosa & A. Cordero
Bio-mathematics	Ezio Venturino & Nico Stollenwerk & Maíra Aguiar
Quasi-Solvable Systems in Quantum Chemistry and Physics	Jerzy Cioslowski
Mathematical Models for Computer Science	Jesús Medina & Manuel Ojeda-Aciego
From clusters to the solid State	Ian Hamilton & Peter Schwerdtfeger
Computational Linear and NonLinear Algebra	P. Alonso
Fixed Point Theory in various abstract spaces and related applications	Regular Session
Computational methods in direct and inverse (systems of) PDE's	Regular Session
Industrial Mathematics	Regular Session
Mathematics in the Information Society	P. Caballero Gil & F.J. Lobillo Borrero & J.A. López Ramos & E. Martínez Moro
Computational methods for fluid flow	Zhenquan Li & Tiejin Wang
Uncertainty and Imprecision Modelling in Decision Making	P. Alonso, H. Bustince, I. Díaz & S. Montes

Acknowledgements

We would like to express our gratitude to the University of Cádiz, especially to the group of Prof Medina of the dept. of Mathematics and professor José Antonio Álvarez Bermejo - Universidad de Almería,

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

- Pedro Alonso, Universidad Oviedo
- Jose Ranilla, Universidad Oviedo
- Raquel Cortina, Universidad Oviedo
- Carmelo Clavero, Universidad Zaragoza
- Higinio Ramos, Universidad Salamanca
- José Antonio Álvarez Bermejo - Universidad de Almería,
- María Eugenia Cornejo Piñero - Universidad de Cádiz,
- Juan Carlos Díaz Moreno - Universidad de Cádiz,
- Eloisa Ramírez Poussa - Universidad de Cádiz,
- M.T. Torres Monteiro Univerisdade Braga Protugal

CMMSE 2016 Plenary Speakers

- Wolfgang Sprößig - TU Bergakademie Freiberg, **Germany**
- Nico Stollenwerk - Lisbon University, **Portugal**
- Humberto Bustince – University Public of Navarra, **Spain**
- Motoko Kotani - Tohoku University, **Japan**

Volume I

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An ecoepidemic predator-prey model with prey vaccination.

Francesca Abbona[1] **and Ezio Venturino**[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino,*
via Carlo Alberto 10, 10123 Torino, Italy

emails: abbona.francesca@gmail.com, ezio.venturino@unito.it

Abstract

In this paper an eco-epidemiological predator-prey model is proposed, in which the presence of a disease in the prey is assumed. Even if the disease does not affect the predators, it can lead to extinction of both populations. Assuming the latter to be a useful resource, the prey vaccination as control measure is introduced in order to protect biodiversity. Boundedness of solutions and equilibria feasibility are obtained. Stability around the different equilibrium points is analyzed through eigenvalues and the Routh-Hurwitz criterion. Simulations are carried out to support the theoretical results.

Key words: vaccination, predator-prey model, population models, epidemics
MSC 2000: AMS codes: 92D25, 92D10, 92D40

1 Introduction

Eco-epidemiology is a branch in mathematical biology which considers simultaneously both the ecological and the epidemiological issues. Ecological populations suffer from various infectious diseases, which have a significant role in regulating population sizes. Mathematical studies of such eco-epidemiological models have explored several unknown aspects of ecological population interactions, see e.g. [6].

In ecosystems, the interaction between the predator and their prey is a nonlinear and rather complex process. The effects of disease-induced mortality or disease-reduced reproduction in regulating natural populations, decreasing their population sizes, reducing their natural fluctuations, or causing destabilization of equilibria into oscillations of the population states have been considered in [3, 4]. The findings of [1] indicate that from both the ecological and the mathematical points of view it is important to study ecological systems

 ISBN: 978-84-608-6082-2

subject to epidemiological factors. The aim of this study is to investigate the impact of a vaccine provided on the infected prey population in an eco-epidemiological mathematical predator-prey model. Costs coming from vaccination strategy can be high. However, it is evident that the impact on the system may be beneficial and thereby regulate and preserve the populations in the system.

The paper is organized as follows. After a brief introduction, In Section 2 basic assumptions are considered and the model is then formulated. Section 3 investigates first of all the model boundedness and consequently the feasible equilibra, the stability of which is hence examined. Finally Section 4 includes numerical simulations, in order to support and illustrate the analytical result achieved.

2 Model Formulation

A mathematical model is proposed, involving three populations: the prey population density N, partitioned among susceptible, infected and vaccinated, $N(t) = S(t) + I(t) + V(t)$, and the predator population $P(t)$. The model assumptions are:
In the absence of the disease, the prey population $N(t)$ grows logistically with intrinsic growth rate r and carrying capacity k.
In the disease presence, prey are partitioned among susceptible $S(t)$ and infected $I(t)$.
Susceptible prey become infected by contact with infected prey, at rate λ; they are also removed by predation at rate h.
Healed infected prey are not immune and return to the susceptible class at rate γ.
Susceptible prey can be vaccinated at rate ν.
Newborns from vaccinated prey are susceptible, whereas infected do not reproduce.
Susceptible, infected and vaccinated experience intraspecific competition and predation.
Predators die at rate n.

Thus the model, in which all parameters are nonnegative, reads

$$\begin{cases} \dfrac{dS}{dt} = rS\Big(1 - \dfrac{S+I+V}{k}\Big) + rV - \lambda IS + \gamma I - \nu S - hSP \\ \dfrac{dI}{dt} = \lambda IS - \mu I - \gamma I - gIP - rI\dfrac{S+I+V}{k} \\ \dfrac{dV}{dt} = \nu S - lVP - rV\dfrac{S+I+V}{k} \\ \dfrac{dP}{dt} = -nP + e\Big(hS + gI + lV\Big)P \end{cases} . \tag{1}$$

The first equation for susceptibles accounts for their reproduction and intraspecific competition, for vaccinated reproduction, the infection process at rate λ, the vaccination at rate ν, disease recovery at rate γ, and predation, at rate h. The second equation for infected includes new recruitments, through successful contact with a susceptible, natural

plus disease-related mortality μ, disease recovery, predation and intraspecific competition. The third equation for the vaccinated class accounts for the new entries, the predation at rate l, and intraspecific competition. Finally the fourth equation for the predators considers their mortality at rate n and the benefit of hunting on the three prey classes, scaled via the conversion coefficient $e < 1$. The Jacobian of (1) is

$$J = \begin{bmatrix} J_{11} & -\frac{r}{k}S - \lambda S + \gamma & -\frac{rS}{k} + r & -hS \\ \lambda I - \frac{rI}{k} & J_{22} & -\frac{rI}{k} & -gI \\ \nu - \frac{rV}{k} & -\frac{rV}{k} & J_{33} & -lV \\ ehP & egP & elP & J_{44} \end{bmatrix} \tag{2}$$

with

$$J_{11} = r - \lambda I - \nu - hP - 2\frac{Sr}{k} - \frac{rI}{k} - \frac{rV}{k}, \quad J_{33} = -lP - \frac{rS}{k} - \frac{rI}{k} - 2\frac{rV}{k},$$
$$J_{22} = \lambda S - \mu - \gamma - gP - \frac{Sr}{k} - 2\frac{rI}{k} - \frac{rV}{k}, \quad J_{44} = -n + e\left(hS + gI + lV\right)$$

3 Model analysis

3.1 Boundedness

To show that the model proposed is well posed, we prove its boundedness. Let $U = S+I+V$ and $W = U + P$. Then

$$\frac{dW}{dt} + \eta W \leq (r+\eta)U - \frac{r}{k}U^2 + (\eta - n)P, \quad \frac{dU}{dt} + \eta U \leq (r+\eta)U - \frac{r}{k}U^2.$$

The right hand side of the second inequality represents a concave parabola Ψ, with maximum value at $\overline{U} = \frac{k(r+\eta)}{2r}$. Hence, $\Psi(\overline{U}) = \frac{k}{4r}(r+\eta)^2$ and we have

$$U' + \eta U \leq \Psi(\overline{U}), \quad W' + \eta W \leq \Psi(\overline{U}) + (\eta - n)P.$$

Since η is arbitrary, we choose it so that $0 < \eta < n$ for which it follows

$$W' \leq \Psi(\overline{U}) - \eta W,$$

giving

$$W \leq \frac{\Psi(\overline{U})}{\eta}(1 - e^{-\eta t}) + W(0)e^{-\eta t}.$$

As $t \to \infty$, we have $W \leq \max\{\Psi(\overline{U})\eta^{-1}, W(0)\} = Z$, implying that the solutions U, i.e. S, I, V, and P are bounded. All the solutions for the model for any $\varepsilon > 0$ are confined in the region:

$$\Gamma = \left\{(S, I, V, P) \in \mathbb{R}^4_+ : W \leq Z + \varepsilon\right\}.$$

 ISBN: 978-84-608-6082-2

3.2 Equilibrium points

In addition to the origin E_0, and coexistence, the remaining equilibria $E_k = (P_k, S_k, I_k, R_k)$ are the following: the predator-and-disease-free point $E_1 = (S_1, 0, V_1, 0)$, the disease-free point $E_2 = (S_2, 0, V_2, P_2)$ and the predator-free point $E_3 = (S_3, I_3, V_3, 0)$. Specifically, we analyse each one of them below.

Equilibrium E_1

For the predator-and-disease-free point respectively the third and first equations in system (1) give the two isoclines:

$$S_1 = \frac{rV^2}{\nu k - rV}, \quad V_1 = \frac{rS + k(\nu - r)}{r(k - S)} S. \tag{3}$$

S_1 has a positive branch raising up from the origin to infinity at the vertical asymptote $\widehat{V} = \nu k r^{-1}$. Instead V_1 has a similar behavior if $r < \nu$, raising up to $\widehat{S} = k$, or from $S_0 = k(1 - \nu r^{-1}) < k = \widehat{S}$ if $r > \nu$. The two curves in the $S - V$ phase subplane meet always in the first quadrant, even when they both go through the origin, since V_1 is convex and S_1^{-1} is concave. The intersection occurs in the box which is the cartesian product of the intervals $0 < V < \nu k r^{-1}$ and $\widehat{m} < S < k$, where we have set $\widehat{m} = \max\{0, S_0\}$. This ensures that both S_1 and V_1 are positive. Therefore the feasible existence of E_1 is always guaranteed.

Equilibrium E_2

For the disease-free point from the third equilibrium equation we have

$$S_2 = \frac{n}{eh} - \frac{l}{h} V, \tag{4}$$

which entails the feasibility condition:

$$0 < V_2 < \widetilde{V} = \frac{n}{el}. \tag{5}$$

The remaining equilibrium populations are the intersection of the following isoclines:

$$P = \varphi_1(V) = \frac{\frac{rl}{kh}\left(\frac{l-h}{h}\right) V^2 - \frac{1}{h}\left(l\nu - r\left(l - 2\frac{nl}{hke} + \frac{n}{ek} - h\right)\right) V - \frac{n}{eh}\left(r - \nu - \frac{rn}{keh}\right)}{lV - \frac{n}{e}} = \frac{\Pi(V)}{l(V - \widetilde{V})}, \tag{6}$$

$$P = \varphi_2(V) = \frac{r(l-h)eV^2 - (\nu lek + rn)V + \nu nk}{kehlV} = \frac{\Gamma(V)}{kehlV}. \tag{7}$$

Study of the curve φ_1

φ_1 is defined for all $V \geq 0$, except for $V = \widetilde{V}$. It intersects the vertical axis at $\widetilde{P}_0 = h^{-1}\left(r\left(1-\nu-n(ekh)^{-1}\right)\right)$, whereas the intersections with the horizontal one, existing if its discriminant Δ_1 is nonnegative, are the roots $V_{2\pm}$ of the parabola $\Pi(V)$ in the numerator. The latter is convex if and only if $l > h$. Let $P_0 = n(eh)^{-1}\left(r\left(1-n(ekh)^{-1}\right)\right) = \Pi(0)$ denote the parabola intersection with the vertical axis. Then $P_0 = -ne^{-1}\widetilde{P}_0$. There are several cases depending on the convexity of Π and on the location of its roots with respect to $\widetilde{V}$. Only in some of them ϕ_1 admits a positive branch, which are discussed below. The ranges for a viable V are specified.
We begin with the cases for $\Delta_1 > 0$ and $l > h$:
[1] $V_- < V_+ \leq 0 < \widetilde{V} \leq V$; here ϕ_1 has a branch coming down on the right of the vertical asymptote at $\widetilde{V}$ to a minimum positive value and then tending to $+\infty$ as $V \to +\infty$.
[2] $V_- \leq 0 \leq V \leq V_+ < \widetilde{V} \leq V$; there is an arc joining $(0, \widetilde{P}_0)$ with $(V_+, 0)$ and further also a branch coming down on the right of the vertical asymptote at $\widetilde{V}$ to a minimum positive value and then tending to $+\infty$ as $V \to +\infty$.
[3] $0 \leq V_- \leq V \leq V_+ < \widetilde{V} < V$; there is a positive concave arc joining $(V_-, 0)$ and $(V_+, 0)$ and a branch coming down on the right of the vertical asymptote at $\widetilde{V}$ to a minimum positive value and then tending to $+\infty$ as $V \to +\infty$.
[4] $V_- < 0 \leq V < \widetilde{V} < V_+ \leq V$; there are two branches raising up, the first one to the vertical asymptote from $(0, \widetilde{P}_0)$, the second one going from the point $(V_+, 0)$ to $+\infty$ as $V \to +\infty$.
[5] $0 < V_- \leq V < \widetilde{V} < V_+ \leq V$; two branches raise up, the first one to the vertical asymptote from $(V_-, 0)$, the second one going from the point $(V_+, 0)$ to $+\infty$ as $V \to +\infty$.
[6] $0 < \widetilde{V} < V \leq V_- < V_+ \leq V$; again there are two branches, one coming down on the right of the vertical asymptote to $(V_-, 0)$, the second one going from the point $(V_+, 0)$ to $+\infty$ as $V \to +\infty$.
Next, the cases for $\Delta_1 > 0$ and $l < h$:
[7] $V_- < V_+ \leq 0 \leq V < \widetilde{V}$; here ϕ_1 has a branch raising up from $(0, \widetilde{P}_0)$ to the vertical asymptote at $\widetilde{V}$.
[8] $V_- \leq 0 \leq V_+ \leq V < \widetilde{V}$; there is an arc raising up from $(V_+, 0)$ to the vertical asymptote at $\widetilde{V}$.
[9] $0 \leq V \leq V_- \leq V_+ \leq V < \widetilde{V}$; there are two branches, the first one coming down from $(0, \widetilde{P}_0)$ to $(0, V_-)$, the second one going from the point $(V_+, 0)$ to the vertical asymptote at $\widetilde{V}$.
[10] $V_- < 0 \leq V_+ \leq V < \widetilde{V}$; there is an arc raising up from $(V_+, 0)$ to the vertical asymptote at $\widetilde{V}$.
[11] $0 \leq V \leq V_- < \widetilde{V} < V \leq V_+$; two branches come down, the first one from $(0, \widetilde{P}_0)$ to $(V_-, 0)$, the second one coming down from the vertical asymptote to the point $(V_+, 0)$.
[12] $0 \leq V < \widetilde{V} \leq V_- \leq V \leq V_+$; one branch raises up from $(0, \widetilde{P}_0)$ to the vertical asymptote

 ISBN: 978-84-608-6082-2

at $\widetilde{V}$, and then a concave arc joining $(V_-, 0)$ with $(V_+, 0)$.
Finally we consider the case for $\Delta_1 < 0$ and $l < h$, as the opposite one cannot arise:
[13] $0 < \widetilde{V} < V$ there is a branch coming down on the right of the vertical asymptote at $\widetilde{V}$ to a minimum positive value and then tending to $+\infty$ as $V \to +\infty$.

Study of the curve φ_2

φ_2 is defined for all positive values of V, with a vertical asymptote at the origin. The intersections $V^{\pm}$ of Γ with the horizontal axis exist depending on its discriminant $\Delta_2 = (rn + \nu lek)^2 - 4rnk\nu(l - h)$.
[A] For $l > h$ and $\Delta_2 < 0$, $\Gamma > 0$ everywhere, so that φ_2 from the vertical asymptote at the origin comes down to a minimum positive value and then raises up to $+\infty$ as $V \to +\infty$.
[B] For $l < h$ and $\Delta_2 > 0$, since $\Gamma(0) > 0$, Γ is positive for $0 < V < V^+$, so that φ_2 comes down from the vertical asymptote at the origin to the value zero at $V = V^+$.
For $l > h$ and $\Delta_2 > 0$, Γ is positive for $V < V^-$ and $V > V^+$, and since $\Gamma(0) > 0$, there are two cases:
[Ca] $V^- < V^+ < 0$, for which φ_2 is convex and comes down from the vertical asymptote at the origin to a minimum positive value and then tends to $+\infty$ as $V \to +\infty$;
[Cb] $0 < V^- < V^+$; in this situation φ_2 comes down from the vertical asymptote at the origin to the point $(V^-, 0)$ and then raises up from $(V^+, 0)$ to $+\infty$.

Study of the intersections of the curves φ_1 and φ_2

We now combine the above situations in order to see whether an intersection between φ_1 and φ_2 is guaranteed. There are other cases in which with further conditions other intersections may occur, but we do not explore them. The intersection exists unconditionally in the following cases. We give also the interval in which the abscissa V_* lies.
For $\Delta_1 > 0$, $\Delta_2 > 0$ and $l > h$: [4Ca] ($V_* < \widetilde{V}$), [5Ca] ($V_- < V_* < \widetilde{V}$), [6Ca] ($\widetilde{V} < V < V_-$).
For $\Delta_1 > 0$, $\Delta_2 < 0$ and $l > h$: [1Cb] ($\max\{V^+, \widetilde{V}\} < V^*$), [2Cb] ($\max\{V^+, \widetilde{V}\} < V^*$), [3Cb] ($\max\{V^+, \widetilde{V}\} < V^*$), [4Cb] ($V^* < \min\{V^-, \widetilde{V}\}$), ; for [5Cb] the intersection exists if $V_- < V^* < \min\{V^-, \widetilde{V}\}$.
For $\Delta_1 > 0$, $\Delta_2 > 0$ and $l < h$: for [7B] and [12B] the location is $V^* < \min\{V^+, \widetilde{V}\}$; in the next three cases the intersection exists conditionally, namely if $V_+ < V^* < \min\{V^+, \widetilde{V}\}$: for [8B], [9B] and [10B]. For [11B] the intersection exists if $\max\{V^+, \widetilde{V}\} < V^* < V^+$.
The case $\Delta_1 < 0$, $\Delta_2 < 0$ and $l > h$ instead does not guarantee an intersection without further stricter conditions, which we do not explore.

Equilibrium E_3

For the predator-free point from the second equation

$$I = \eta_1(S, V) = \left(\frac{\lambda k}{r} - 1\right) S - V - \frac{k}{r}(\mu + \gamma). \tag{8}$$

From (8) I is nonnegative if:

$$V_3 < \left(\frac{\lambda k}{r} - 1\right) S_3 - \frac{k}{r}(\mu + \gamma). \tag{9}$$

Substituting (8) into the other equations, two functions are obtained, depending on S:

$$V = \eta_2(S) = \frac{\nu}{\lambda S - (\mu + \gamma)} S, \tag{10}$$

$$V = \eta_3(S) = \frac{\lambda^2 k S^2 - \{r(r - \nu + \mu) + \lambda k[\gamma + r(\mu + \gamma)]\} S + k\gamma(\mu + \gamma)}{r\,(\lambda S + r - \gamma)} = \frac{\Omega(S)}{r\lambda(S - \overline{S})}. \tag{11}$$

Study of the curve η_2

The function η_2 goes through the origin, it has a vertical asymptote at $\widehat{S} = (\mu + \gamma)\lambda^{-1}$ but it is positive only when $S > \widehat{S}$. It is always decreasing, having an always negative derivative

$$S' = -\nu \frac{\mu + \gamma}{[\lambda S - (\mu + \gamma)]^2}. \tag{12}$$

Therefore there is only one possibility:
[I] η_2 it comes down from the asymptote at $S = \widehat{S}$ to the horizontal asymptote $\widehat{V} = \nu\lambda^{-1}$.

Study of the curve η_3

The isocline η_3 is feasible for $S \geq 0$, it has a vertical asymptote at $\overline{S} = (\gamma - r)\lambda^{-1}$. It crosses the vertical axis at the height $V_0 = k\gamma(\mu+\gamma)[r(r-\gamma)]^{-1}$. The parabola Ω is convex, with positive height at the origin, $\Omega(0) = k\gamma(\mu + \gamma) > 0$.
The roots of Ω are complex if its discriminant is negative, $\Delta_3 < 0$. Then for η_3 there is:
(a) a positive branch from $(0, V_0)$ tending to $+\infty$ as $S \to +\infty$ if $\overline{S} < 0$, i.e. $r > \gamma$;
(b) a positive branch coming down from the right of the vertical asymptote at $S = \overline{S}$ to a minimum and then tending to $+\infty$ as $S \to +\infty$ if $\overline{S} > 0$, i.e. $r < \gamma$.
For $\Delta_3 > 0$ since $\Omega(0) > 0$, both roots have the same sign. For negative roots we have:
(c) $\overline{S} < S_- < S_+ < 0$, (d) $S_- < \overline{S} < S_+ < 0$, (e) $S_- < S_+ < \overline{S} < 0$, all imply that for η_3 there is a positive branch from $(0, V_0)$ tending to $+\infty$ as $S \to +\infty$;
(f) $S_- < S_+ < 0 < \overline{S}$ the positive branch of η_3 comes down from the right of the vertical asymptote at $\overline{S}$ to a positive minimum and then raises up to $+\infty$ as $S \to +\infty$.
When the roots are positive instead we find:
(g) $\overline{S} < 0 < S_- < S_+$ gives to two positive branches for η_3, one coming down from $(0, V_0)$ to $(S_-, 0)$, the other one raising up from $(S_+, 0)$ to $+\infty$ as S grows;
(h) $0 < \overline{S} < S_- < S_+$ gives to two positive branches for η_3, one coming down from the right

 ISBN: 978-84-608-6082-2

of the vertical asymptote at $\overline{S}$ to $(S_-, 0)$, the other one from $(S_+, 0)$ to $+\infty$ as S grows;
(i) $0 < S_- < \overline{S} < S_+$ gives to two positive branches for η_3, one raising up from $(S_-, 0)$ to the vertical asymptote, the other one raising up from $(S_+, 0)$ to $+\infty$ as S grows;
(j) $0 < S_- < S_+ < \overline{S}$ gives two positive branches for η_3, a concave one joining $(S_-, 0)$ with $(S_+, 0)$, the other one coming down from the right of the vertical asymptote to a positive minimum and then raising up to $+\infty$ as $S \to +\infty$.

Study of the intersections of the curves η_2 and η_3

Combining these cases, a feasible intersection always occurs at the abscissa S^* for:
For [Ia], [Ic], [Id], [Ie] the intersection lies in $\widehat{S} < S^*$; for [Ig], [Ih] it lies in $S_+ < S^*$;
[Ib], [If], [Ij], all give an intersection $\widehat{S} < S^*$ if $\overline{S} < \widehat{S}$;
[Ii] gives an intersection $S_+ < S^*$ if $\overline{S} < \widehat{S}$ and a second one $\widehat{S} < S^* < \overline{S}$ if $\widehat{S} < \overline{S}$.

Equilibrium E_3 is feasible if η_2 and η_3 intersect as discussed above and (9) holds.

Finally, the coexistence equilibrium will be analyzed numerically.

3.3 Stability analysis

Equilibrium E_0

The Jacobian here shows two negative eigenvalues, $\Lambda_1 = -(\mu + \gamma)$ and $\Lambda_2 = -n$. The other ones are the roots of the equation $\Lambda^2 - \mathrm{tr}(\overline{J}(E_0))\Lambda + \det(\overline{J}(E_0)) = 0$ coming from a minor $\overline{J}$ of the Jacobian, where $\mathrm{tr}(\overline{J}(E_0))$ and $\det(\overline{J}(E_0))$ provide the stability condition $\mathrm{tr}(\overline{J}) = r - \nu < 0$ but the remaining one cannot be satisfied, $\det(\overline{J}) = -r\nu < 0$ so that E_0 is always unstable.

Equilibrium E_1

At E_1 we find

$$\Lambda_1 = J_{22}|_{E_1} = \lambda S - \frac{r}{k}S - \frac{r}{k}V - \mu - \gamma, \quad \Lambda_2 = J_{44}|_{E_1} = -n + ehS + elV$$

The other two eigenvalues do not affect the result. In fact, the Routh-Hurwitz conditions on the remaining 2 by 2 minor $\widetilde{J}$ are satisfied,

$$\mathrm{tr}(\widetilde{J}(E_1)) = -\frac{r}{k}(S+V) - r\frac{V}{S} - \nu\frac{S}{V} < 0, \quad \det(\widetilde{J}(E_1)) = \frac{r}{k}\left(r\frac{V^2}{S} + \nu\frac{S^2}{V} + rV + \nu S\right) > 0.$$

Thus the stability conditions are:

$$\frac{h}{l}\left(\frac{n}{eh} - S\right) > V > \frac{\lambda k - r}{r}S - \frac{k}{r}(\mu + \gamma) \tag{13}$$

The second condition (13) is exactly the opposite of (9), thereby showing a transcritical bifurcation between E_3 and E_1, while comparing the first one with (4), it follows that whenever E_2 is feasible, E_1 must be unstable. Thus, when equilibrium E_1 is stable, equilibria

E_2 and E_3 do not exist. On the other hand, if at least one of them is feasible, then E_1 is unstable.

Equilibrium E_2

Here one eigenvalue is explicitly obtained, $\Lambda_2 = J_{22}|_{E_2} = \lambda S - (\mu+\gamma) - gP - rk^{-1}(S+V)$. The characteristic polynomial is the cubic

$$a_0\Lambda^3 + a_1\Lambda^2 + a_2\Lambda + a_3 = 0$$

with

$$a_1 = \frac{r}{k}(S+V) + \nu\frac{S}{V} + r\frac{S}{V}, \quad a_3 = hl(\nu S + rV) + \left(rl^2\frac{V^2}{S} + \nu h^2\frac{S^2}{V}\right) + \frac{r}{k}(l-h)^2SV,$$
$$a_0 = 1, \quad a_2 = 2\frac{r^2}{k^2}SV + el^2VP + 2r\nu + \frac{r}{k}\left(r\frac{V^2}{S} + \nu\frac{S^2}{V}\right) - \frac{r}{k}(\nu S + rV) + eh^2SP.$$

Stability occurs when the Routh-Hurwitz conditions hold, but since $a_1 > 0$ and $a_3 > 0$, we need only to require the last one, as well as the negativity of the eigenvalue $\Lambda_2 < 0$:

$$a_1a_2 > a_3, \quad \lambda S_2 < (\mu+\gamma) + gP_2 + \frac{r}{k}(S+V). \tag{14}$$

Equilibrium E_3

At $E_3 = (S_3, I_3, V_3, 0)$, again one eigenvalue is $\Lambda_4 = J_{44}|_{E_3} = -n + e\,(hS + gI + lV)$. Using the Routh-Hurwitz criterion for the remaining cubic characteristic equation $\sum_{i=0}^{3} b_i\Lambda^i$, with $b_0 = 1$

$$\begin{aligned}
b_1 &= -\left(\frac{r}{k}I + r\frac{V}{S} - \left(\frac{r}{k}(S+V) + \gamma\frac{I}{S} + \nu\frac{S}{V}\right)\right)\\
b_2 &= \frac{r}{k}(\gamma+r)\frac{IV}{S} + \frac{r\nu(S+V)S}{kV} + \nu r + \lambda\gamma I +\\
&\quad - \left(\frac{r^2(S+V)I}{k^2} + \frac{r(rV^2+\gamma I^2)}{kS} + I\left(\lambda^2 S + \frac{\nu\gamma}{V} + \frac{r\gamma}{k}\right)\right)\\
b_3 &= I\left(\frac{r^3V^2}{k^2S} + \frac{r\nu S}{kV} + \frac{r\lambda^2S^2}{k} + \nu\lambda^2\frac{S^2}{V} + \frac{r\nu\gamma}{k}\right) +\\
&\quad - I\left(\frac{r^2\gamma IV}{k^2S} + \frac{r\gamma\lambda V}{k} + \lambda\nu\gamma\frac{S}{V} + \frac{r^3}{k^3}S^2 + \frac{r^2\nu S^2}{k^2V} + \frac{r\lambda\nu S}{k}\right)
\end{aligned}$$

Stability is achieved when $\Lambda_4 < 0$, and the Routh-Hurwitz conditions hold, i.e. explicitly

$$\begin{gathered}
hS + gI + lV < \tfrac{n}{e}, \quad b_1b_2 > b_3, \quad \tfrac{r}{k}I + r\tfrac{V}{S} < \tfrac{r}{k}(S+V) + \gamma\tfrac{I}{S} + \nu\tfrac{S}{V}, \\
\tfrac{r}{k}\left(\tfrac{r^2}{k}\tfrac{V^2}{S} + \nu\tfrac{S}{V} + \lambda^2S^2 + \nu\gamma\right) + \nu\lambda\gamma\tfrac{S^2}{V} > \tfrac{r}{k}\left(\tfrac{r}{k}\gamma\tfrac{IV}{S} + \tfrac{r^2}{k^2}S^2 + \tfrac{r}{k}\nu\tfrac{S^2}{V} + \gamma\lambda V + \lambda\nu S\right).
\end{gathered} \tag{15}$$

 ISBN: 978-84-608-6082-2

4 Simulations

The simulations, implemented with the Matlab solver ode45, are run for selected sets of parameter values, in part obtained from the literature and in part taken as hypothetical. They are summarized in Table1.

Parameter	Value	Explanation	Source
r	11.2	intrinsic growth rate	[5],[6]
K	10000	carrying capacity	hypothetical
λ		susceptible to infected rate of conversion	simulated
μ	0.4	infected mortality due to disease	[5],[6]
n	0.3	predators mortality	[7]
ν		vaccination rate	simulated
h	0.1	predation of susceptible	hypothetical
g	0.3	predation of infected	hypothetical
l	0.1	predation of vaccinated	hypothetical
γ		infected to susceptible conversion rate	simulated
e	0.25	conversion factor of prey into new predators	[5],[6]

Table 1: Model parameters

Figure 4, left frame, represents the endemic state: the four populations coexist. In this way the infection is not eradicated. Coexistence is feasible with the parameters choice: $\lambda = 5$, $\nu = 0.3$, $\gamma = 11$. In this way only a very small portion of the susceptible population is treated with vaccine. Moreover, it is assumed that the disease spreads easily, but also infected recover at high rate. The same situation is reproduced with $\lambda = 9$: infection is assumed to have a higher incidence with the same portion of vaccinated as in the previous situation, in the right frame of Figure 4. The equilibrium E_2 is stably attained for the following choice of parameters: $\lambda = 1.2$, $\nu = 10$, $\gamma = 3$. Initial conditions are $S = 100$, $I = 500$, $M = 10$, $N = 50$. Infected prey become quickly extinct and susceptible prey, vaccinated and predators show decaying oscillations before they reach the stable values, 4 left frame. Even on the assumption of having a greater number of contacts $\lambda = 5$ between susceptible and infected prey, with the same rate of vaccination equilibrium E_2 is achieved, see right frame of Figure 4.

5 Conclusions

In this paper, a nonlinear predator-prey ecoepidemic model is proposed and analyzed in order to study the effect of a vaccine on the prey population suffering from a disease.

Apart from the extinction point, other four equilibria have been shown to be feasible

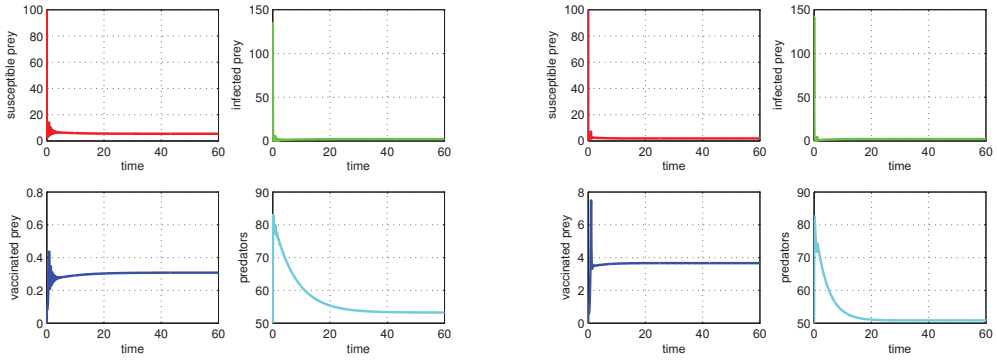


Figure 1: Susceptible, infected, vaccinated and predators around coexistence with parameters given in Table 1 and with $\lambda = 5$, $\nu = 0.3$, $\gamma = 11$ (left) and with $\lambda = 5$, $\nu = 0.3$, $\gamma = 11$ (right).

and two different situations can arise. In some conditions, the predator-free equilibrium point with endemic disease is a possible case. When this equilibrium becomes unstable, the system settles to the infection-and-predator-free point. Here only susceptible prey are found, which is not a good result from the biodiversity point of view. On the other hand, the two populations can coexist with the endemic disease. Thus in this case the vaccine cannot eradicate the disease. But in suitable conditions disease eradication is possible, showing that vaccine is effective. In this case biodiversity is preserved. The numerical simulations show that the populations can coexist when the disease is highly virulent, with large incidence rate and when vaccinations are not administered or alternatively implemented at a too low rate. Increasing the latter, good results are obtained: it is possible to overcome the illness and to preserve the different populations in the environment. The theoretical analysis allows to compare different types of vaccination control measures and their effect on the populations global evolution.

Acknowledgments

Work partially supported by projects “Metodi numerici in teoria delle popolazioni” and “Metodi numerici nelle scienze applicate” of Dipartimento di Matematica Univ. di Torino.

References

[1] M. Haque and E. Venturino, *Increase of the prey may decrease the healthy predator population in presence of a disease in the predator*, Hermis, **7:** 38-59 (2006).

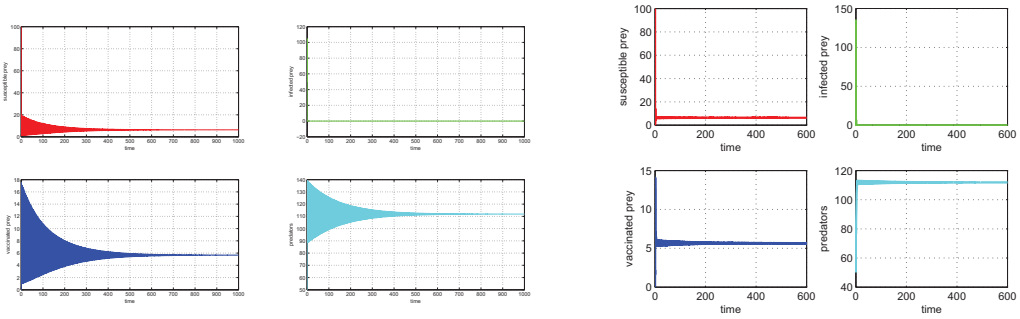


Figure 2: Susceptible, infected, vaccinated prey and predators around disease-free equilibrium with $\lambda = 1.2$, $\nu = 10$, $\gamma = 3$ (left) and with $\lambda = 5$, $\nu = 10$, $\gamma = 3$ (right).

[2] M. Haque and E. Venturino, *The role of transmissible diseases in Holling Tanner predator prey model*, Journal of Theoretical Population Biology, **70**(3) 273-288 (2006).

[3] H. W. Hethcote, Z. Ma, and S. Liao, *Effects of quarantine in six endemic models for infectious diseases*, Mathematical Biosciences, **180:** 141-160 (2002).

[4] H. W. Hethcote,W.Wang, and Z.Ma, *A predator-prey model with infected prey*, Theoretical Population Biology, **66:** 259-268 (2004).

[5] Hugo A., Massawe Estomih S. and Makinde Oluwole D., *An eco-epidemiological mathematical model with treatment and disease infection in both prey and predator population.*, Journal of Ecology and The Natural Environment, **4**(10) 266-279 (2012).

[6] Mukhopadhyay B., Bhattacharyya R. , *Dynamics of a delay-diffusion prey-predator model with disease in the prey.*, J. Appl. Math. and Computing **17**(1-2) 361-377 (2005).

[7] P. Tengaa, D. O. Makinde , E. S. Massawe, *An Eco-Epidemiological Model with Nonlinear Incidence and Infective Prey Treatment Class*, International Journal of Modern Trends in Engineering and Research (IJMTER), **2**(11) 308-329 (2015).

[8] Venturino E., *The influence of diseases on Lotka Volterra systems.*, Rocky Mountain Journal of Mathematics, **24:** 381-402 (1994).

[9] Venturino E., *Epidemics in predator-prey models: disease in prey.*, in: Mathematical Population dynamics, Analysis of heterogeneity 1, (Eds.:O. Arino, D. Axelrod, M., Kimmel, Langlais, M.), Wuertz, Winnipeg, Canada, 381-393 (1995).

[10] Venturino E., *Simple Metaecoepidemic Models*, Bull Math Biol, **73:** 917-950 (2011).

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Hub-directed multigraphs and arrowhead matrices

J. Abderramán Marrero[1], Juan Núñez[2] and María Trinidad Villar[2]

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

[2] *Department of Geometry and Topology, University of Seville (Spain)*

emails: jc.abderraman@upm.es, jnvaldes@us.es, villar@us.es

Abstract

In this paper we use the arrowhead matrices as a tool to study Graph Theory, more precisely, an interesting class of directed multigraphs, the hub-directed multigraphs. We associate the arrowhead matrices with the adjacency matrices of a class of directed multigraph and we obtain new properties of the second objects by using properties of the first ones. The hub-directed multigraphs with potential use in applications are also defined. As main result, we show that a hub-directed multigraph $G(H^*)$ with adjacency matrix H^* is a dominant hub-directed multigraph if and only if $H^* = CE$, where C is the adjacency matrix of another directed multigraph, and E is the adjacency matrix of a particular elementary dominant hub-directed pseudo-graph. Another decomposition of its Gram (arrowhead) matrix $A = (H^*)^T H^*$ is also given.

Key words: arrowhead matrices, hub-matrices, directed multigraphs
MSC 2000: 65F15; 05C50; 05C20.

1 Introduction

Using relationships between different knowledge fields leads to one of the most stimulating research in Mathematics. Indeed, with this approach, new alternative techniques raise to improve well-known theories or find new ones. As an example of it, the reader can consult the research started in [4, 5], where a mapping between Lie algebras and combinatorial structures was introduced in order to translate properties of Lie algebras into the language of Graph Theory and vice versa, research which was later extended to Leibniz algebras in [6]. In an analogous way, several works concerning certain connection between the evolution algebras and Graph Theory can be found in [10, 11].

In this paper, we are studying and analyzing the link between Graph Theory and arrowhead matrices. We use also hub-matrices in this study.

Regarding Graph Theory, the study of its properties and applications is currently running in a high level due to its very widespread use as a tool to solve many important problems in other disciplines. Indeed, graphs (and more particularly, trees) have been essential previously to study several properties on semi-simple Lie algebras due to its role in determining the Dynkin diagrams associated to such algebras [13].

With respect to arrowhead matrices, there exists a current and increasing interest in the use of those of large order in the theory of the hub matrices, due to their useful applications in networks, wireless communication, and the world wide web (see [9, 14] and the references therein, for instance).

The main goal of this paper is to obtain an association between arrowhead matrices and graphs, particularly directed multigraphs. The aim is to make easier the study of each of them by dealing with their associated object. Properties of each of them could be later translated into the language of the other.

The structure of the paper is as follows: First, we recall some preliminaries on arrowhead matrices, hub-matrices and directed multigraphs. Section 3 is devoted to the study of hub-directed multigraphs, which allows us to set the relationship between them and the arrowhead matrices. By using its link we obtain that a hub-directed multigraph with adjacency matrix H^* is a dominant hub-directed multigraph if and only if $H^* = CE$, where C is the adjacency matrix of another directed multigraph and E the adjacency matrix of a particular elementary dominant hub-directed pseudo-graph. Another decomposition of its Gram (arrowhead) matrix $A = (H^*)^T H^*$ is also given. In this way, we have given some steps towards a better knowledge of arrowhead matrices by means of Graph Theory. In Section 4 some applications and extensions of this research are outlined briefly.

2 Preliminaries

We recall some basic features regarding arrowhead matrices [8, 12, 15], hub matrices [9], and directed multigraphs [2, 3, 7]. In this work, only the finite case will be considered.

2.1 Arrowhead matrices

A matrix $A \in \mathbb{K}^{n\times n}$, with $\mathbb{K}$ an arbitrary real or complex field, is *arrowhead* if A is of the form

$$A = \left(\begin{array}{c|cccc} b_0 & c_1 & c_2 & \cdots & c_{n-1} \\ \hline a_1 & b_1 & 0 & \cdots & 0 \\ a_2 & 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & 0 & 0 & \cdots & b_{n-1} \end{array}\right), \tag{1}$$

 ISBN: 978-84-608-6082-2

in short $A = \{a_i, b_i, c_i\}_n$. If the entries a_i, $i = 1, 2 \ldots, n-1$ are zero, A is called an *upper triangular arrowhead matrix*. Alternatively, if the entries $c_i = 0$, $i = 1, 2 \ldots, n-1$ are zero, A is a *lower triangular arrowhead matrix*.

Triangular arrowhead matrices have interesting properties. For instance, if A, B are $n \times n$ upper triangular arrowhead matrices, and $\lambda \in \mathbb{K}$, then $A + B$, λA, and AB are also upper triangular arrowhead matrices. If A is a nonsingular lower triangular arrowhead matrix, its matrix inverse $A^{-1} = (d_{ij})_{i,j=1}^n$ has the entries

$$d_{ij} = \begin{cases} 1/b_i, & \text{if } i = j, \\ -a_i/b_i, & \text{for the entries of the first column } j \neq i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The same properties are verified for upper triangular arrowhead matrices. The proofs of these sentences are trivial.

The determinant, $|A|$, of an arrowhead matrix A as given in (1), with $b_j \neq 0$ for $j = 1, 2, \ldots, n-1$, can be obtained easily by expanding $|A|$ successively by its first row or column.

$$|A| = b_0 \prod_{i=1}^{n-1} b_i - \sum_{i=i}^{n-1} a_i c_i \prod_{\substack{j=2 \\ j\neq i}}^{n-1} b_j = \left[b_0 - \sum_{i=i}^{n-1} \frac{a_i c_i}{b_i} \right] \prod_{j=1}^{n-1} b_j = u \prod_{j=1}^{n-1} b_j, \tag{3}$$

where the value u,

$$u = b_0 - \sum_{i=1}^{n-1} \frac{a_i c_i}{b_i}, \tag{4}$$

is either zero or nonzero depending on whether A is singular or nonsingular, respectively.

However in the symmetric eigenvalue problem [12, 15] and also in real-world applications [9], the term arrowhead is most commonly used for the symmetric arrowhead matrices. From here on, we take this convention by assuming $c_i = b_i$, and denoting $A = \{a_i, b_i\}_n$ without loss of generality. Some known results [1] are adapted here to the symmetric case.

Theorem 1 ([1]). *Every arrowhead matrix $A = \{a_i, b_i\}_n$, with $b_i \neq 0$, $i = 1, 2, ..., n-1$, has a UDU^T factorization of the form*

$$A = \begin{pmatrix} 1 & \frac{a_1}{b_1} & \frac{a_2}{b_2} & \cdots & \frac{a_{n-1}}{b_{n-1}} \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} \begin{pmatrix} u & & & & \\ & b_1 & & & \\ & & b_2 & & \\ & & & \ddots & \\ & & & & b_{n-1} \end{pmatrix} \begin{pmatrix} 1 & & & & \\ \frac{a_1}{b_1} & 1 & & & \\ \frac{a_2}{b_2} & & 1 & & \\ \vdots & & & \ddots & \\ \frac{a_{n-1}}{b_{n-1}} & & & & 1 \end{pmatrix}, \tag{5}$$

where the value $u = b_0 - \sum_{i=1}^{n-1} a_i^2/b_i$.

 ISBN: 978-84-608-6082-2

Note that the matrix D is invertible for A nonsingular. Thus the matrix inverse A^{-1} can be factored trivially.

Corollary 1 ([1]). *The inverse A^{-1} of a nonsingular arrowhead matrix $A = \{a_i, b_i\}_n$ satisfying Theorem 1, with $u \neq 0$, has a triangular factorization of the form, $A^{-1} = LD^{-1}L^T$, where $L = \left(U^T\right)^{-1}$.*

From Corollary 1 we can obtain a simple decomposition of A^{-1}, which can be computed in linear time.

Corollary 2 ([1]). *The matrix inverse A^{-1} of a nonsingular arrowhead matrix $A = \{a_i, b_i\}_n$ satisfying Theorem 1, with $u \neq 0$, can be decomposed as a symmetric rank-one perturbation of a singular diagonal matrix,*

$$A^{-1} = \begin{pmatrix} 0 & & & & \\ & \frac{1}{b_1} & & & \\ & & \frac{1}{b_2} & & \\ & & & \ddots & \\ & & & & \frac{1}{b_{n-1}} \end{pmatrix} + \frac{1}{u} \begin{pmatrix} 1 \\ \frac{-a_1}{b_1} \\ \frac{-a_2}{b_2} \\ \vdots \\ \frac{-a_{n-1}}{b_{n-1}} \end{pmatrix} \begin{pmatrix} 1 & \frac{-a_1}{b_1} & \frac{-a_2}{b_2} & \cdots & \frac{-a_{n-1}}{b_{n-1}} \end{pmatrix}. \quad (6)$$

2.2 Hub-matrices

A hub-matrix $H \in \mathbb{R}^{n\times m}$ is a matrix whose first column (called hub-column) has an Euclidean norm greater than or equal to the one of any other column, and the usual scalar product of the hub-column with any other column is nonzero. In addition, these remaining $n-1$ columns (called nonhub-columns) are orthogonal each other with respect to the scalar product. In other words, if H is a hub-matrix then its associated Gram matrix $A = H^T H$ is an arrowhead matrix; see Theorem 1 from [9]. The fact that H is a hub-matrix is sufficient to claim that its Gram matrix $A = H^T H$ is an arrowhead matrix, but the hub condition is not necessary. For its possible extension to the infinite case, we have taken the first column as the hub-column, instead of the last one defined in [9].

An example of a hub-matrix and its associated Gram matrix is the following

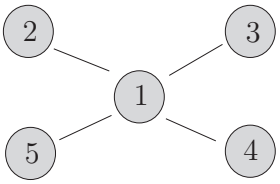
$$H = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 3 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}, \qquad A = H^T H = \begin{pmatrix} 15 & 1 & 2 & 5 \\ 1 & 2 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 5 & 0 & 0 & 5 \end{pmatrix}. \quad (7)$$

2.3 Directed multigraphs

Recall that a *simple graph* is a pair $G = (V, E)$ where V is a finite non-empty set of elements, called *vertices*, and E is a set of 2-subsets of elements of V, called *edges*. The *adjacency*

matrix of a simple graph with vertex set $V = \{v_1, \ldots, v_n\}$ is an $n \times n$ matrix, where the element in the i-th row and j-th column is 1 if there exists an edge from vertex v_i to vertex v_j (for $1 \leq i, j \leq n$) and 0 otherwise. The adjacency matrix of any simple graph is symmetric and binary.

The simple graph denoted by K_1 has vertex set $V = \{v_1\}$ and edge set $E = \emptyset$. The simple, noncomplete, graph denoted by $K_{1,n}$ has vertex set $V = \{v_1, \ldots, v_{n+1}\}$ and edge set $E = \{\{v_1, v_i\}$, for $i = 2, \ldots n+1\}$. Notice that $K_{1,n}$ displays the *star topology* (in fact, it is named the *star graph*), where the non-hub vertices have a unique edge connected with the hub-vertex $\{v_1\}$. The adjacency matrix of $K_{1,n}$ is a traceless zero-one (binary) arrowhead matrix.



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

Figure 1: The simple graph $K_{1,4}$ and its adjacency matrix.

A *directed multigraph* is a pair $G = (V, D)$, where V is a finite non-empty set of elements called *vertices* and D is a set of ordered pairs of elements of V, called *directed edges*, admitting *loops* which are edges with both vertices equal. When multiple edges are not allowed, we refer to these graphs as *directed pseudo-graphs*, and they can be considered as a particular case of directed multigraphs. When neither multiple edges nor *loops* are allowed, we have a *simple directed graph*, *digraph* shortly.

The *adjacency matrix* of a directed multigraph with vertex set $V = \{v_1, \ldots, v_n\}$ is an $n \times n$ matrix with nonnegative entries. When there are α edges from vertex v_i to vertex v_j, for $1 \leq i, j \leq n$, the (ij)-entry is α, a nonnegative integer. For a directed pseudo-graph, its adjacency matrix is binary since $\alpha = 1$ or $\alpha = 0$.

A *morphism* between two directed multigraphs, $G_1 = (V_1, D_1)$ and $G_2 = (V_2, D_2)$, is a map $\phi : V_1 \to V_2$ such that if $(u, v) \in D_1$, then $(\phi(u), \phi(v)) \in D_2$. An *isomorphism* between directed multigraphs is a bijective morphism and its converse is also morphism.

Given a square matrix A of order n with nonnegative entries, there is an only directed multigraph associated to A, up to isomorphism. Conversely, given a directed multigraph G, its adjacency matrix may be not unique. In general, it depends on the labelling of the vertex set of G, $V = \{v_1, \ldots, v_n\}$. However, if A and A' are two adjacency matrices of G, the matrices A and A' are cogredient, i.e. there is a permutation matrix P verifying: $A' = P^{-1}AP = P^TAP$.

 ISBN: 978-84-608-6082-2

3 Hub-directed multigraphs

In the general case, the adjacency matrix of a directed multigraph maybe not a hub-matrix. Nevertheless, given a square hub-matrix H of order n, we are interested in considering the directed multigraph of n vertices whose adjacency matrix is precisely the hub-matrix H.

$$B = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

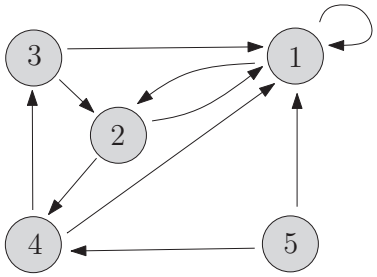


Figure 2: Hub-directed pseudo-graph whose adjacency matrix is B.

When the hub-matrix H is of order $n \times m$ $(n \neq m)$, it is possible to add some rows or columns of zeros in order to obtain a square matrix denoted H^*. Therefore, H^* can be considered as an adjacency matrix of a new directed multigraph, the hub-directed multigraph. We introduce the following

Definition 1. *Given a hub-matrix H of order $n \times m$, the hub-directed multigraph, $G(H)$ of $max\{n, m\}$ vertices is the graph whose adjacency matrix is precisely H^* defined as follows:*

- *If $max\{n, m\} = n$, H^* is the matrix H with $n - m$ columns of zeros added on the right.*
- *If $max\{n, m\} = m$, H^* is the matrix H with $m - n$ rows of zeros added on the bottom.*
- *In both cases, $G(H)$ is called the hub-directed multigraph associated to H.*

Next, we consider the transpose of the hub-matrix B given in Figure 2 that is also a square matrix, but it is not a hub-matrix. Its associated directed pseudo-graph is depicted in Figure 3.

Let us illustrate an example of an arrowhead matrix, which is the Gram matrix, B^TB, of the hub-matrix B. The associated directed multigraph to B^TB, Figure 4, has multiple edges and it has the simple graph with *star topology* $K_{1,3} \cup K_1$ as subjacent graph, which is the simple graph obtained by suppressing the loops and considering only the simple edges $\{\{1, 2\}, \{1, 3\}, \{1, 4\}\}$; see e.g. Figure 1 for $K_{1,4}$.

$$B^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

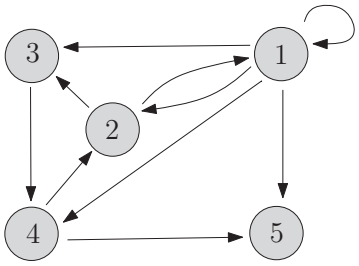


Figure 3: Directed pseudo-graph associated to the matrix B^T.

$$B^T B = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 1 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Note that this property can be generalized to the case of any hub-matrix H as follows

Proposition 1. *Let H be a hub-matrix of order $n \times m$. The arrowhead matrix $A = H^T H$ is the adjacency matrix of a directed multigraph whose subjacent graph is $K_{1,n}$, which displays the* star topology*, with possibly some isolated vertices.*

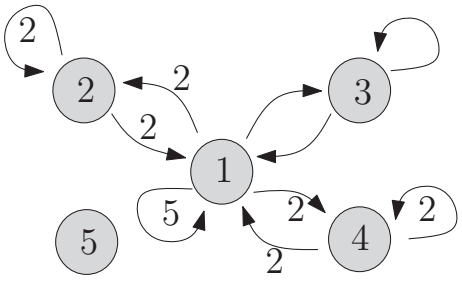


Figure 4: Directed multigraph associated to the arrowhead matrix $B^T B$.

3.1 Dominant hub-directed multigraph

We are interested in hub-directed multigraphs that can be factored as the composition of another directed multigraph by a simpler hub-directed pseudo-graph.

 ISBN: 978-84-608-6082-2

Definition 2. *A hub-directed multigraph whose adjacency matrix satisfies* $< \boldsymbol{v}_1, \boldsymbol{v}_j > \geq < \boldsymbol{v}_j, \boldsymbol{v}_j >$*, for every nonzero column* $\boldsymbol{v}_j$*, with the usual scalar product* $<,>$*, is called a dominant hub-directed multigraph.*

Notice that the hub-directed pseudo-graph from Figure 2 is a dominant hub-directed pseudo-graph. However, if we zeroing the $(3,1)$-entry of the adjacency matrix B of Figure 3, because now $< \mathbf{v}_1, \mathbf{v}_2 > = \ 1 \ < \ 2 \ = < \mathbf{v}_2, \mathbf{v}_2 >$, the new hub-directed pseudo-graph is not dominant.

Definition 3. *An elementary dominant hub-directed pseudo-graph is a dominant hub-directed pseudo-graph whose adjacency matrix* H^* *satisfies* $< \boldsymbol{v}_1, \boldsymbol{v}_j > = \ 1 \ = < \boldsymbol{v}_j, \boldsymbol{v}_j >$*, for every nonzero binary column* $\boldsymbol{v}_j$*, with* $j \neq 1$*.*

The adjacency matrix of an elementary dominant hub-directed pseudo-graph characterizes the dominant hub-directed multigraph. We now introduce a new concept.

Definition 4. *The* $n \times n$ *square hub-matrix* E *which has one-entries on its hub-column and for* $j = 2, ..., n$*, the nonhub-columns of* E *are the same as those of the* $n \times n$ *unit matrix is called* elementary hub-matrix of order n*.*

Lemma 1. *The matrix* H^* *satisfies* $H^* = CE$*, where* C *coincides with* H^* *with the exception that the first column of* C *is the hub-column of* H^* *minus the nonhub-colums of* H^* *and* E *is the elementary hub-matrix.*

Remark 1. *Notice that the matrix* C *has a directed multigraph associated, while* E *is associated with an elementary dominant hub-directed pseudo-graph. Observe also that the elementary hub-matrix of order* n*,* E*, is the adjacency matrix of a particular elementary dominant hub-directed pseudo-graph with the star topology. See, for instance, the graph on the right of Figure 5.*

Definition 5. *A decomposition given by lemma 1,* $H^* = CE$*, is called* $C-E$ decomposition*.*

Theorem 2. *A hub-directed multigraph* $G(H^*)$ *with adjacency matrix* H^* *is a dominant hub-directed multigraph if and only* H^* *admits a* $C - E$ *decomposition, where* E *is the elementary hub-matrix and* C *is the adjacency matrix of another directed multigraph associated to* $G(H^*)$*.*

The proof is trivial by observing that $C = H^*E^{-1}$. As an illustration, we observe the $C - E$ decomposition of the adjacency matrix of a dominant hub-directed multigraph. Figure 5 shows the associated graphs.

$$H^* = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 7 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 4 & 0 & 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 4 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

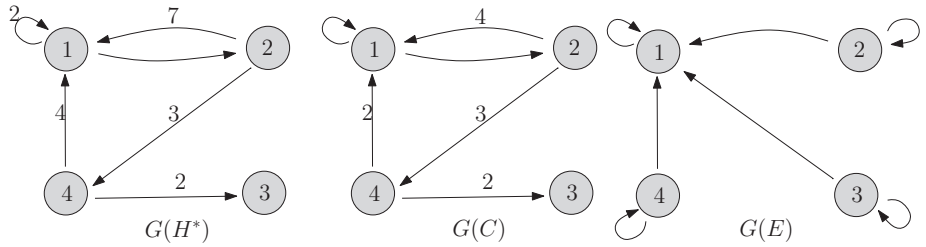


Figure 5: $C - E$ decomposition of the hub-directed multigraph with adjacency matrix H^*.

Note that, in the particular case of hub-directed pseudo-graphs, the following result holds.

Corollary 3. *A hub-directed pseudo-graph $G(H^*)$ with adjacency matrix H^* is a dominant hub-directed pseudo-graph if and only if H^* can be factored in the form $H^* = CE$, with C the adjacency matrix of another directed pseudo-graph and E the elementary hub-matrix.*

More precisely, the adjacency matrix C is the same than H^*, but the first column of C is a zero column.

As an illustration, the decomposition of the adjacency matrix B of the dominant hub-directed pseudo-graph from Figure 2 is $B = CE$,

$$B = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Notice as the factorization (5) for an arrowhead matrix $A = B^T B$ yields $A = E^T DE$, where B is the adjacency matrix of a dominant hub-directed pseudo-graph, Corollary 3. Here the diagonal matrix $D = C^T C$. Moreover, as an immediate consequence of Theorems 1 and 2, we can obtain the following results regarding dominant hub-directed multigraphs and arrowhead matrices. It can be applied as a particular case to dominant hub-directed pseudo-graph.

Lemma 2. *The arrowhead matrix $A = (H^*)^T H^*$ satisfies $A = DE$, where D coincides with A with the exception that the first column of the matrix D is the same as A but the new first column is the first column of A minus the remaining columns of A and E is the elementary hub-matrix.*

Definition 6. *A decomposition given by LEMMA 2, $A = DE$ is called* $D-E$ decomposition.

Theorem 3. *A hub-directed multigraph $G(H^*)$ with adjacency matrix H^* is a dominant hub-directed multigraph if and only if its Gram (arrowhead) matrix $A = (H^*)^T H^*$ admits a $D - E$ decomposition, where E is the elementary hub-matrix and D is the adjacency matrix of another directed multigraph.*

For example, the Gram matrix A related with the adjacency matrix B of the dominant hub-directed pseudo-graph from Figure 2 yields $A = DE$,

$$A = \begin{pmatrix} 5 & 2 & 1 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 & 1 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The matrix D from Theorem 3 is related with degrees and loops of the vertices of the factored directed multigraph with adjacency matrix A. Indeed, the (ij)-entry $(j \neq 1)$ of D gives us the number of edges from vertex v_i to vertex v_j (including loops) of the directed multigraph associated to the Gram matrix A. The $(1j)$-entry $(j \neq 1)$ of D gives us the number of edges from vertex v_j to the vertex v_1 minus the number of loops of the vertex v_j. The (11)-entry gives the number of loops of the vertex v_1 minus the number of (no loops) edges starting from v_1.

4 Conclusions

The hub-directed multigraph, in particular the dominant hub-directed multigraph, has been introduced and characterized through its adjacency matrix H^* and also its Gram (arrowhead) matrix $A = (H^*)^T H^*$. The results obtained can be of use in applications [9, 14], e.g. in the study of the gap and the ratio of second to first eigenvalue of a dominant hub-matrix H with respect to a nondominant one, which is a matrix rearrangement of H. The same can be used for the last to first eigenvalue. Also it can be applied in the performance of hub-networks, with star topology, with respect to other (ring, hybrid) network topologies. Of both theoretical and applied interest can be the characterization of the hub-directed multigraph isomorphism.

Acknowledgements

The authors gratefully acknowledge financial support by the Spanish Ministerio de Ciencia e Innovación and Junta de Andalucía via grants No. MTM2013-40455-P FEDER, MTM2015-65397-P and No. FQM-326 (J.Núñez.), FQM-164 (M.T.Villar), respectively.

References

[1] J. Abderramán Marrero, V. Tomeo, *Infinite invertible arrowhead matrices and applications*, Proceedings of the 15th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2015 Vol **1** (2015) 1-11.

[2] B. Bollobás, *Graphs Theory*, Springer Verlag, New York, USA, 1979.

[3] J. A. Bondy and U. S. R. Murty , *Graph Theory with Applications*, Springer, New York, USA, 2010.

[4] A. Carriazo, L.M. Fernández, J. Núñez, *Combinatorial structures associated with Lie algebras of finite dimension*, Linear Algebra Appl. 389 (2004), 43-61.

[5] M. Ceballos, J. Núñez, A. F. Tenorio, *Study of Lie algebras by using combinatorial structures*, Linear Algebra Appl. 436 (2012), 349–363.

[6] M. Ceballos, J. Núñez, A. F. Tenorio, *Low-dimensional Leibniz algebras and combinatorial structures*. Mathematics and Computers in Simulation (2016). In press.

[7] F. Harary, *Graph Theory*, Addison Wesley, Reading, Mass., USA, 1969.

[8] N. Jakovcevic Stor, I. Slapnicar, J. L. Barlow, *Accurate eigenvalues decomposition of real symmeric arrowhead and applications*, Linear Algebra Appl. **464** (2015) 62-89.

[9] H. T. Kung, B. W. Sutter, *A hub matrix theory and applictions to wireless communications*, EURASIP J. Adv. Signal Process. (2007) Article ID 13659, 8 pages.

[10] J. Núñez, M.L. Rodríguez-Arévalo and M.T. Villar, *Certain particular families of graphicable algebras*, Applied Mathematics and Computation **246**:1 (2014) 416-425.

[11] J. Núñez, M. Silvero and M. T. Villar *Mathematical tools for the future: Graph Theory and graphicable algebras*, Applied Mathematics and Computation, **219**:11 (2013) 6113-6125.

[12] D. P. O'leary, G. W. Stewart *Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices*, J. Comput. Phys. **90** (1990) 497505.

[13] J.P. Serre, *Algèbres de Lie Semi-Simples Complexes*, Benjamin Inc., New York, 1996.

[14] L. Shen, B. W. Sutter, *Bounds for eigenvalues of arrowhead matrices and their applications to hub matrices and wireless communications*, EURASIP J. Adv. Sgnal Process. (2009) Article ID 379402, 12 pages.

[15] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, New York, U.S.A. 1965.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Fast algorithms for solving general k-tridiagonal matrix linear equations

Jesús Abderramán Marrero[1]

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

emails: jc.abderraman@upm.es

Abstract

To generalize the current inversion procedures for $n \times n$ nonsingular k-tridiagonal matrices $T_n^{(k)}$, a fast and accurate numerical solver is proposed for the matrix equation $T_n^{(k)} X = B$. The solution is evaluated in $\mathcal{O}(n^2)$ time at most, using Givens reduction and adapted back substitution. In particular, the matrix inverse of $T_n^{(k)}$ is computed in $\mathcal{O}(n^2/k)$ time. The solution of the k-tridiagonal vector linear equation is also evaluated in $\mathcal{O}(n)$ time.

Key words: Fast sequential algorithms, Givens reduction, k-tridiagonal matrix linear equation.
MSC 2000: 15B99, 65F99.

1 Extended abstract

The tridiagonal matrices appear frequently in the solution of initial and boundary value problems in differential equations that modeling many real-world applications. It gave rises to design special algorithms for inverting such matrices; see e.g. [1, 2] and references therein. A generalization of the tridiagonal matrix is the k-tridiagonal matrix, GNkT matrix shortly for the nonsingular case, with entries

$$\left[T_n^{(k)}\right]_{ij} = \begin{cases} d_i, & \text{if } i = j \\ a_i, & \text{if } j - i = k; \\ b_i, & \text{if } i - j = k; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The inversion of the GNkT matrices is of current interest; see e.g. [3, 4, 5]. It motivates our proposal of a specific numerical and sequential procedure for solving the general GNkT matrix linear equation,

$$T_n^{(k)} X = B, \quad \text{where } X,\ B \in \mathcal{M}_{n\times m}(\mathbb{R}), \tag{2}$$

with arbitrary $m \in \mathbb{N}$. The matrix X is the solution to be evaluated. If $m = n$ and $B = I_n$, the solution is $X = \left(T_n^{(k)}\right)^{-1}$, the matrix inverse of $T_n^{(k)}$. Specially, we have the GNkT vector linear equation,

$$T_n^{(k)} x = b; \quad T_n^{(k)} \in \mathbb{R}^{n\times n}; \quad x, b \in \mathbb{R}^n, \tag{3}$$

We focus without loss of generality on the vector linear equation (3). In [3, 4] the matrix inverse of $T_n^{(k)}$ was evaluated using recurrence relations and, if necessary, a symbolic variable. However, procedures based on recurrence relations are frequently unreliable [6]. Overflow or underflow can appear. Also, the symbolic procedure has resulted slow, even unfeasible, for large order matrices. In [4, 5] was suggested the division in k proper tridiagonal matrices obtained from the GNkT matrix $T_n^{(k)}$ using elementary modular arithmetic. The tridiagonal matrices were inverted separately. Then, the entries of such inverses were reallocated in the adequate positions of $\left(T_n^{(k)}\right)^{-1}$ to complete the matrix inverse. Although of possible interest in parallel procedures, managing modular arithmetic, matrix division, symbolic variable, and matrix reallocation reduces the speed of direct and sequential algorithms.

To avoid the explicit computation of the matrix inverse (unless it coincides with the solution of (2)) and the previous drawbacks, Givens reduction of $T_n^{(k)}$ to triangular form is proposed, which preserves sparsity. Although complex Givens rotations are available, we handle real GNkT matrices. Thus the simpler system, equivalent to (3), yields

$$Ux = v; \quad U = Q^T T_n^{(k)} \in \mathbb{R}^{n\times n}; \quad x, v = Q^T b, \in \mathbb{R}^n, \tag{4}$$

where the new coefficient matrix U is an upper triangular matrix, with a well-defined zero band-pattern, with only one (for $n \le 2k$) or two ($2k < n$) nonzero bands above the main diagonal. In the orthogonal reduction, the usual procedure for computing the coefficients of the Givens rotations can be used. The particular structure of U is exploited using an adapted back substitution scheme for solving the system (4) in linear time.

The simple procedure is summarized in the following algorithm. Its extension to the GNkT matrix linear equation (2) is immediate.

Algorithm 1
Input: The $n \times n$ GNkT matrix $T_n^{(k)}$ and the vector b from (3) **Output:** An approximation of the vector solution x
Step 1: Computing U and v from (4) 1.a) Givens rotations to obtain U zeroing the k-th subdiagonal of $T_n^{(k)}$. 1.b) The same Givens rotations on b to obtain v.
Step 2: Solving the equivalent determined vector equation (4). 2.a) Computing x from (4) using back subtitution.

Degree	500	1500	2000	5000	7500
mldivide	$3.97e^{-3}$	$7.64e^{-2}$	$1.68e^{-1}$	$2.10e^{-0}$	$6.54e^{-0}$
Algorithm 2	$4.01e^{-3}$	$1.70e^{-2}$	$2.59e^{-2}$	$1.17e^{-1}$	$2.41e^{-1}$

Table 1: Mean elapsed times (in seconds and after 100 trials) for computing the vector solution of Example 1.

Degree	500	1500	2000	5000	7500
mldivide	$6,31e^{-16}$	$1,09e^{-15}$	$1,22e^{-15}$	$1.89e^{-15}$	$2.44e^{-15}$
Algorithm 2	$1,24e^{-15}$	$2,19e^{-15}$	$2,65e^{-15}$	$4.14e^{-15}$	$4.94e^{-15}$

Table 2: Residual for the outcomes of Example 1.

1.1 Numerical examples

Two current examples are used for the numerical comparison of the Algorithm 1, and its slight adaptation for computing the matrix inverse, called Algorithm 2, with respect to the advanced mldivide built-in solver of *Matlab®'s* package ($R2015a$). For a fair comparison, we take *Matlab®'s* variable MaxNumCompThreads = 1. The trials have been done in a CPU Intel® Core $i7$ 4770 of 3.40 GHz and 8 GB of RAM.

Example 1. *The performance of the Algorithm* 1 *with respect that of the mldivide solver is compared by solving Equation* (3)*. The* $n \times n$ *GNkT matrix* $T_n^{(2)}$ *is as given for* $n = 7$ *in [5], Section* 3*. That is,* $\left[T_n^{(k)}\right]_{ii} = 8$ *for* i *odd (*1 *for* i *even);* $\left[T_n^{(k)}\right]_{i,i-k} = 2$ *for* $i-k$ *odd (*1 *for* $i-k$ *even);* $\left[T_n^{(k)}\right]_{j-k,j} = -2$ *for* $j-k$ *odd (*1 *for* $j-k$ *even);* $\left[T_n^{(k)}\right]_{ij} = 0$*, otherwise.*

The vector $b \in \mathbb{R}^n$ *has entries* $b(i) = (-1)^i i/(n+i)$*. In Table 1, the mean elapsed times are compared. In Table 2, the residual* $||T_n^{(2)}\hat{x} - b||_F$ *(Frobenius norm) are checked.*

Example 2. *We compare in Table 3 the accuracy of the Algorithm* 2 *with respect that of Algorithm* 3.2 *from [4], KTRINV Algorithm given in [3], and the usual mldivide solver, by evaluating* $||T_n^{(k)}X - I_n||_F/||I_n||_F$*, the residual norm. The GNkT matrix to be inverted is the diagonally dominant matrix* $T_n^{(k)}$*, with* $k = 2$*, given in [4], Example* 4.3*, with entries*

$$\left[T_n^{(k)}\right]_{ij} = \begin{cases} 8.5 + (-1)^i 0.5, & \text{if } i = j; \\ -1.5 + (-1)^j 0.5, & \text{if } i - j = 2; \\ -3.5 + (-1)^i 0.5, & \text{if } j - i = 2; \\ 0, & \text{otherwise.} \end{cases}$$

 ISBN: 978-84-608-6082-2

Degree	**Jia-Li*, [4]**	**KTRINV*, [3]**	**mldivide**	**Algorithm 2**
1000	$1.8010e^{-16}$	$1.8053e^{-16}$	$1.8106e^{-16}$	$1,0760e^{-16}$
1500	$1.8120e^{-16}$	$1.8171e^{-16}$	$1.8213e^{-16}$	$1,0191e^{-16}$
2000	$1.8175e^{-16}$	$1.8229e^{-16}$	$1.8259e^{-16}$	$9.891e^{-17}$
2500	$1.8208e^{-16}$	$1.8264e^{-16}$	$1.8294e^{-16}$	$9.707e^{-17}$
3000	$1.8230e^{-16}$	$1.8288e^{-16}$	$1.8312e^{-16}$	$9.582e^{-17}$

Table 3: Residual norm $||T_n^{(2)}X - I_n||_F/||I_n||_F$ supply for the algorithms using the matrix of Example 2. The outcomes of algorithms marked with * were obtained from [4], Table 2.

Degree	1000	1500	2000	2500	3000
mldivide	$8.38e^{-2}$	$2.565e^{-1}$	$5.892e^{-1}$	$1.1089e^{0}$	$1.8733e^{0}$
Algorithm 2	$1.12e^{-2}$	$2.30e^{-2}$	$4.00e^{-2}$	$5.96e^{-2}$	$8.49e^{-2}$

Table 4: Mean elapsed times (in seconds and after 100 trials) for computing the matrix inverse of the GNkT matrix $T_n^{(k)}$ of Example 2 with $k = n/2$.

The inverse is the solution of Equation (2)*, taking* $B = I_n$*. For* $k = n/2$*, the (mean) times elapsed by Algorithm* 2 *with respect to that of the mldivide solver, in the computation of the inverse of* $T_n^{(k)}$*, are detailed in Table 4.*

References

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

[2] J. Abderramán Marrero, M. Rachidi, , *A note on representations for the inverses of tridiagonal matrices*, Linear Multilinear Algebra **61** (2013) 1181–1191.

[3] M. E. A. El-Mikkawy, F. Atlan, *A novel algorithm for inverting a general k-tridiagonal matrix*, Appl. Math. Lett. **32** (2014) 41–47.

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

[5] C. M. da Fonseca, F. Yilmaz, *Some comments on k-tridiagonal matrices: Determinant, spectra, and inversion*, Appl. Math. Comput. **270** (2015) 644–647.

[6] N. J. Higham, *Accuracy and stability of numerical algorithms*, 2^{nd} Ed. SIAM, Philadelphia, PA, USA, 2002.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A class of tests for the two-sample problem for count data based on the empirical probability generating function

M. Virtudes Alba-Fernández[1], Apostolos Batsidis[2], M. Dolores Jiménez-Gamero[3] and Pedro Jodrá[4]

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

[2] *Department of Mathematics, University of Ioannina, Greece*

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

[4] *Department of Statistic Methods, University of Zaragoza, Spain*

emails: mvalba@ujaen.es, abatsidis@uoi.gr, dolores@us.es, pjodra@unizar.es

Abstract

A class of tests for the two-sample problem for count data whose test statistic is an L_2-norm of the difference between the empirical probability generating functions is considered. The null distribution of the test statistic is unknown, so some approximations are investigated. Specifically, the bootstrap, permutation and weighted bootstrap estimators are considered. All of them provide consistent estimators. A simulation study analyzes the performance of these approximations for small and moderate sample sizes. This study also includes a comparison with other two-sample tests based on comparing the empirical characteristic functions associated to the samples.

Key words: two-sample problem, count data, probability generating function, simulation

1 Introduction

The problem of testing whether two samples come from the same population is a statistical issue of great interest and many different approaches have been proposed to deal with it. One of them is related to the use of the characteristic function (CF) and its empirical counterpart (ECF) by means of an L_2-norm between both ECFs. Since the resultant test statistic is not distribution free, Meintanis [2] suggested that it could be approximated by

means of permutation and bootstrap procedures (see Alba-Fernández et al. [3] for a sound justification). Jiménez-Gamero et al. [4] studied the use of a weighted bootstrap estimator (in the sense of Burke [5]). The latter estimator is showed to have the same asymptotic properties than the former ones, but it is more efficient from the computational point of view.

This paper studies the two-sample problem for count data. When dealing with this sort of data, Nakamura and Pérez-Abreu [1] argue in favor of using inferential methods based on the empirical probability generating function (EPGF). The motivation for using the PGF is that it is usually much simpler than the corresponding probability mass function (PMF), fully characterizes the distribution and possesses convenient features not shared by the characteristic or moment generating function such as being a real valued continuous analytic function which always exists in the range $[0,1]^d$, where d is the dimension of the random vector under study.

Motivated by Nakamura and Pérez-Abreu [1], a class of tests based on the L_2-norm between the EPGFs associated with both samples is considered. The limiting distribution of the test statistic under the null hypothesis is derived. It is not distribution free, so some approximations, such as bootstrap, permutation and weighted bootstrap, are studied.

Because the tests based on the EPGFs and those based on ECFs have similar asymptotic properties, a simulation study is carried out in order to investigate the performance of both approaches for small or moderate sample sizes. As anticipated by Nakamura and Pérez-Abreu [1], we found that the test based on comparing the EPGFs behaves better than that based on the ECFs.

2 The two-sample problem

Let X and Y be two random vectors taking values in $\mathbb{N}_0^d$, for some fixed $d \in \mathbb{N}$, with cumulative distribution functions (CDF) F_X and F_Y, respectively. Let us consider the problem of testing for the equality of both distributions. The null hypothesis is stated as

$$H_0 : F_X(x) = F_Y(x), \quad \forall x \in \mathbb{N}_0^d \quad \Longleftrightarrow \quad C_X(t) = C_Y(t), \quad \forall t \in \mathbb{R}^d, \tag{1}$$

where C_X and C_Y are the probability generating functions (pgf) of X and Y, respectively. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent random samples from X and Y, with sizes n and m, respectively. For testing (1), we consider the following test function

$$\Phi_{n,m} = \begin{cases} 1, & \text{if } D_{n,m} \geq d_{n,m,\alpha}, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where

$$D_{n,m} = \int \left| C_{X,n}(t) - C_{Y,m}(t) \right|^2 w(t) dt,$$

w is a probability density function defined on $[0,1]^d$, $C_{X,n}(t)$ and $C_{Y,m}(t)$ denote the empirical probability generating functions (egpf) associated with the samples,

$$C_{X,n}(t) = \frac{1}{n}\sum_{j=1}^{n} t^{X_j}, \quad C_{Y,m}(t) = \frac{1}{m}\sum_{l=1}^{m} t^{Y_l},$$

and $d_{n,m,\alpha}$ is the $1-\alpha$ percentile of the null distribution of $D_{n,m}$.

To decide when to reject H_0, that is, to calculate $d_{n,m,\alpha}$ or, equivalently, to calculate the p-value of the observed value of the test statistic, we need to know the null distribution of $D_{n,m}$. In general, this task is quite difficult, so one has to approximate it. Here, we consider some approximations to the null distribution of $D_{n,m}$: the bootstrap (B), the permutation (P) and the weighted bootstrap (WB). It can be shown that these methods provide consistent estimations of the null distribution. For saving space the theoretical results are omitted. Next section summarizes some simulation results.

3 Simulation study

This section empirically investigates the finite sample size properties of the proposed approximations by means of several numerical experiments. All computations in this paper have been performed using programs written in the R language (http://www.cran.r-project.org).

First, we study the goodness of the B, the P and the WB approximations to the null distribution of the test statistic $D_{n,m}$. For doing this, we generated two independent samples with equal sample sizes $n = m = 20$, from a univariate Poisson distribution with mean $\lambda = 2$, $P(2)$; the approximation of the $p-$value (for P, B and WB approximations), say $\hat{p}$, was calculated by considering 1000 replications. As weight function $w(t)$ several options are possible. We considered the following two: the uniform weight $U(0,1)$, and the Beta weight $\beta(2,2)$. The experiment was repeated 1000 times and the fraction of $\hat{p}$s less than or equal to 0.05, i.e. the estimated type I error probability for $\alpha = 0.05$, was calculated. For the WB approximation, as multipliers data from a univariate standard normal distribution were generated. As competitors, in the simulation experiment, we have also included the test statistic for testing (1) similar to that presented here but based on the ECFs. Based on Alba-Fernández et al. [3] as weight function of the ECFs test the uniform and the standard normal distribution was considered. The whole experiment was repeated for $n = m = 50$ and the results are displayed in Table 1. Looking at this table we see that the estimated type I error probabilities are quite close to the nominal values; all approximations –B, P and WB– show similar results.

Next, we compared the approximations of the null distribution of $D_{n,m}$ in terms of the power, and we also compared them with the test based on the ECFs. Here, we have only included the estimated power when the first sample is coming from a Poisson distribution while the second one from the discrete Lindley (DL) distribution introduced by Gómez and

 ISBN: 978-84-608-6082-2

Table 1: Estimated type I errors for the univariate Poisson with mean 2, $P(2)$.

				PGF		
Sample size	*Uniform weight*			*Beta weight*		
$n = m$	B	P	WB	B	P	WB
20	0.053	0.055	0.062	0.056	0.052	0.064
50	0.059	0.055	0.065	0.057	0.057	0.064
100	0.045	0.047	0.050	0.047	0.046	0.048
				ECF		
Sample size	*Uniform weight*			*Normal weight*		
$n = m$	B	P	WB	B	P	WB
20	0.056	0.061	0.068	0.057	0.061	0.071
50	0.047	0.048	0.057	0.046	0.050	0.051
100	0.046	0.046	0.050	0.048	0.050	0.052

Calderín [6]. This last distribution was obtained by discretizing the continuous model of the Lindley distribution. The resultant model is over-dispersed and competitive with the Poisson distribution. Figure 1 displays the PMF of the Poisson and DL laws considered for the power study. The estimated powers are shown in Table 2. From the results given in Table 2 it can be said that the test based on the EPGFs provides better results (in term of the power) than those obtained by using the ECFs. Other alternative distributions were also considered and the obtained results follow similar conclusions.

Acknowledgements

This work has been partially supported by grants: MTM2014-55966-P of the Spanish Ministry of Economy and Competitiveness (M. Dolores Jiménez-Gamero) and CMT2015-68276-R of the Spanish Ministry of Economy and Competitiveness (M. Virtudes Alba-Fernández and Pedro Jodrá).

References

[1] M. Nakamura, and V. Pérez-Abreu, *Empirical probability generating function. An overview*, Insurance: Mathematics and Economics. **12** (1993) 287–295.

[2] S. G. Meintanis, *Permutation tests for homogeneity based on the empirical characteristic function*, J. Nonparametr.Stat. bf 17 (2005) 265–275.

Table 2: Estimated power ($\alpha = 0.05$). Poisson versus Discrete Lindley laws.

	PGF					
	$Uniform\ weight$			$Beta\ weight$		
$P(0.75)$ vs. $DL(0.25)$	B	P	WB	B	P	WB
$n = m = 20$	0.345	0.347	0.393	0.297	0.285	0.330
$n = m = 50$	0.752	0.750	0.769	0.638	0.639	0.654
$P(0.5)$ vs. $DL(0.1)$						
$n = m = 20$	0.333	0.321	0.352	0.245	0.238	0.277
$n = m = 50$	0.676	0.664	0.682	0.494	0.493	0.507
	ECF					
	$Uniform\ weight$			$Normal\ weight$		
$P(0.75)$ vs. $DL(0.25)$	B	P	WB	B	P	WB
$n = m = 20$	0.123	0.122	0.148	0.115	0.118	0.138
$n = m = 50$	0.282	0.289	0.295	0.289	0.291	0.302
$P(0.5)$ vs. $DL(0.1)$						
$n = m = 20$	0.095	0.092	0.114	0.148	0.149	0.182
$n = m = 50$	0.127	0.134	0.136	0.402	0.404	0.420

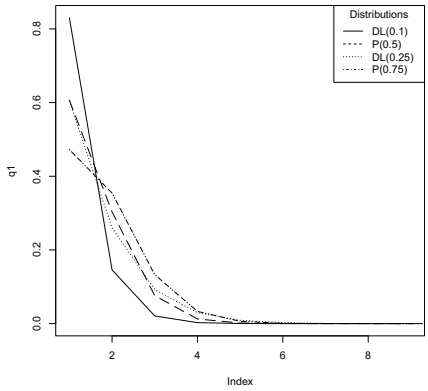


Figure 1: PMF of Poisson distributions and Discrete Lindley alternatives.

[3] V. Alba-Fernández, M. D. Jiménez-Gamero, J. Muñoz-García, *A test for the two-sample problem based on empirical characteristic functions*, Comput. Stat. Data Anal. **52** (2008) 3730–3748.

[4] Jiménez-Gamero, M. D., Alba-Fernńdez, M. V., Jodrá, P., Barranco-Chamorro, I., *Fast tests for the two-sample problem based on the empirical characteristic function*, Submitted (2015)

[5] M. D. Burke, *Multivariate tests-of-fit and uniform characteristic function*, Ann. Probab. Lett. **46** (2000) 13–20.

[6] E. Gómez-Deniz, E. Calderín-Ojeda, *The discrete Lindley distribution: properties and applications*, J. Stat. Comput. and Simul. **81** (2011) 1405–1416.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Critical Sections and Software Transactional Memory Comparison in the Context of a TLS Runtime Library

Sergio Aldea[1], Diego R. Llanos[1] and Arturo Gonzalez-Escribano[1]

[1] *Departamento de Informática, Universidad de Valladolid*

emails: sergio@infor.uva.es, diego@infor.uva.es, arturo@infor.uva.es

Abstract

Transactional Memory (TM) is a technique that aims to mitigate the performance losses that are inherent to the serialization of accesses in critical sections. Some studies have shown that the use of TM may lead to performance improvements, despite the existence of management overheads. However, the relative performance of TM, with respect to classical critical sections management depends greatly on the actual percentage of times that the same data is handled simultaneously by two transactions. In this paper, we compare the relative performance of the critical sections provided by OpenMP with respect to two Software Transactional Memory (STM) implementations. These three methods are used to manage concurrent data accesses in ATLaS, a software-based, Thread-Level Speculation (TLS) system. The complexity of this application makes it extremely difficult to predict whether two transactions may conflict or not, and how many times the transactions will be executed. Our experimental results show that the STM solutions only deliver a performance comparable to OpenMP when there are almost no conflicts. In any other case, their performance losses make OpenMP the best alternative to manage critical sections.

Key words: Software Transactional Memory, STM, Thread-Level Speculation, TLS, OpenMP, ATLaS

1 Introduction

Current multicore processors offer an opportunity to speed up the computation of sequential applications. To exploit these parallel technologies, the software needs to be parallelized, that is, transformed in order to correctly distribute the work among different threads. This process usually involves synchronizing the accesses to certain memory areas that are

shared by the concurrent threads, with the aim of avoiding potential data races. This synchronization is usually performed by using critical sections that protect shared memory structures.

To simplify this process, parallel programming models such as OpenMP [1] offer compiler directives, not only to parallelize the code, but also to synchronize accesses and define and manage critical sections. Despite their simplicity, these solutions present a problem: Critical sections introduce performance losses, not only because they serialize the code, but also because of the cost associated to locking management.

Software Transactional Memory (STM) [2] arises as a possible solution to the first problem, allowing programmers to transform critical sections in transactions that are concurrently and atomically executed. This is based on the optimistic assumption that the code inside the transaction will access to different locations of the shared memory being protected. In these cases, accesses are carried out concurrently. If this is not the case, conflictive transactions should be rolled back and executed one at a time.

Works such as [3] have shown that STM can outperform OpenMP critical sections, despite the relatively high overheads of STM. However, the relative performance of STM versus OpenMP critical sections is highly dependent on the running profile of each particular application. Different patterns of accesses to the same critical section may lead to different performance figures.

This paper compares the OpenMP critical sections approach with two STM libraries, using them to handle the critical sections that appears in the runtime library of ATLaS [4], a state-of-the-art, software-based Thread Level Speculation (TLS) system. Our goal is to study the relative performance of both approaches when managing concurrent accesses in such a complex piece of code.

The rest of this paper is structured as follows: Section 2 briefly describes the fundamentals of software TLS. Section 3 details how our TLS runtime library handles the speculative execution of a source code, and how critical data structures are protected to ensure correctness. Section 4 describes how this protection can be ensured using OpenMP and two different STM libraries. Section 5 shows the performance results obtained by each OpenMP and the STM libraries considered. Finally, Sect. 6 concludes this paper.

2 Thread-Level Speculation in a Nutshell

Speculative parallelization (SP), also called Thread-Level Speculation (TLS) or Optimistic Parallelization [5, 6, 7, 8, 9, 10, 11], is a technique that allows the parallel execution of fragments of code (typically blocks of iterations of a loop) without the need for a compile-time analysis, which guarantees that the fragments do not present data dependences between them. Instead, TLS solutions assume that the loop can be optimistically executed in parallel, and rely on a runtime monitor to ensure that no dependence violations appear. TLS

solutions can be implemented in software or hardware. From here on, we will focus on software-based TLS.

A dependence violation appears when a given thread generates a datum that has already been consumed by a thread executing a subsequent set of iterations with respect to the original sequential order. In this case, the results calculated so far by the successor (called the offending thread) are not valid and should be discarded. Early proposals [5, 6] stop the parallel execution and restart the loop serially. Other proposals stop the offending thread and all its successors, re-executing them in parallel [7, 8, 9, 10].

Figure 1 shows an example of thread-level speculation. The figure represents four threads executing one out of four consecutive iterations, and the sequence of events that occurs when the loop is executed in parallel. All threads access certain data elements from the `SV` vector. If the values of `x` are not known at compile time, the compiler is not able to ensure that accesses to the `SV` structure do not lead to dependence violations when executing them in parallel. However, the indexes of the data elements being accessed are known at runtime, so dependence violations can be detected and corrected while the program is running.

Speculative parallelization works as follows. If the programmer labels the SV vector as speculative, the code should be instrumented at compile time to monitor at runtime that all uses of `SV` follow sequential semantics. At runtime, each thread maintains a version copy of the elements of the `SV` vector being accessed. All read operations to `SV` are replaced by a function that performs a *speculative load*. This function obtains the most up-to-date value of the element being accessed. This operation is called *forwarding*. If a predecessor (that is, a thread executing an earlier iteration) has already read or written that element then the value is forwarded (as Thread 2 does in Fig. 1). If not, then the function obtains the value from the main copy of the speculative data structure (as Thread 3 does in the figure).

Regarding modifications to the speculative data structure, all write operations are replaced at compile time by a *speculative store* function. This function writes the datum in the version copy of the current processor, and ensures that no thread executing a subsequent iteration has already consumed an outdated value for this structure element, a situation called "dependence violation". If such a violation is detected, the offending thread and its successors are stopped and restarted, in a so-called *squash* operation.

If no dependence violation arises for a given thread, it should *commit* all the data stored in its version copy to the main copy of the speculative structure. Note that commits should be done in order, to ensure that the most up-to-date values are stored. After performing the commit operation, a thread can assign itself a new iteration or block of iterations to continue the parallel work.

 ISBN: 978-84-608-6082-2

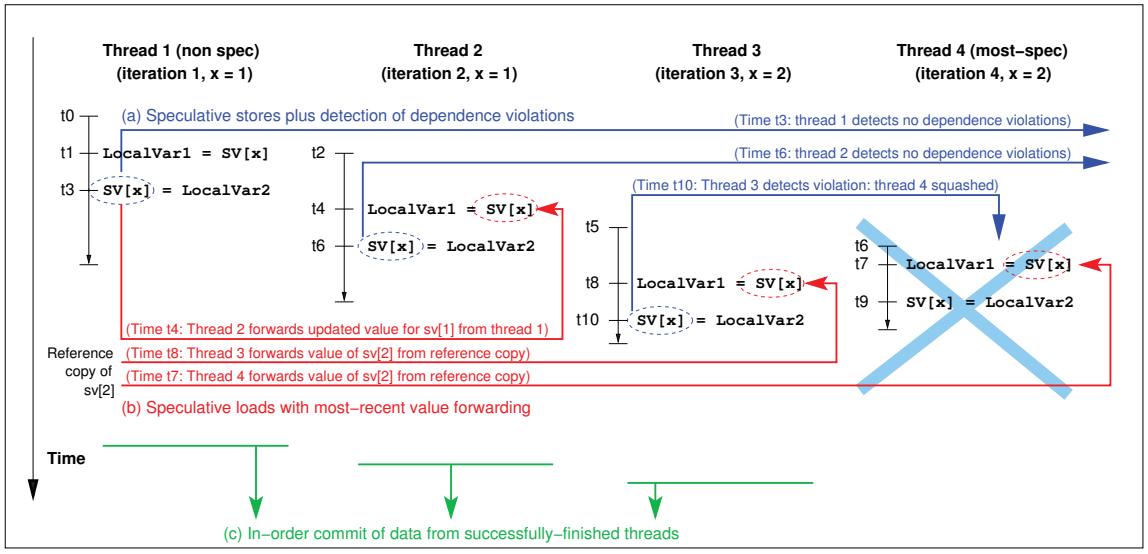


Figure 1: Example of speculative execution of a loop and summary of operations carried out by a runtime TLS library.

3 The ATLaS framework and runtime library

We have developed an extension to OpenMP that incorporates Thread-Level Speculation support. The ATLaS framework [4] allows any loop to be executed in parallel without the need of a prior dependence analysis. This is done by defining a new OpenMP variable classification clause, namely speculative. If the user is unsure about whether the access in parallel to a variable or structure inside a given loop may lead to a dependence violation, he/she may simply classify it as speculative, instead of labeling it as private or shared. In this case, the source code is instrumented at compile time to add TLS execution support. This is done with the help of a GCC compiler plugin [12] that transforms the code, inserting calls to the ATLaS TLS runtime library. When running in parallel, the runtime library ensures that all the accesses to all data elements classified as speculative follows sequential semantics.

The ATLaS runtime library [13] supports all the operations described in the previous section. It follows the design principles of the speculative parallelization library developed by Cintra and Llanos [7], with several improvements that allow, for example, the speculative parallelization of loops that use pointer arithmetic or complex data structures.

One of the key advantages of this library over previous designs is that the ATLaS runtime library is almost free of critical sections. The only critical section needed is the one that manages the data structure that maintains the assignment of chunks of iterations

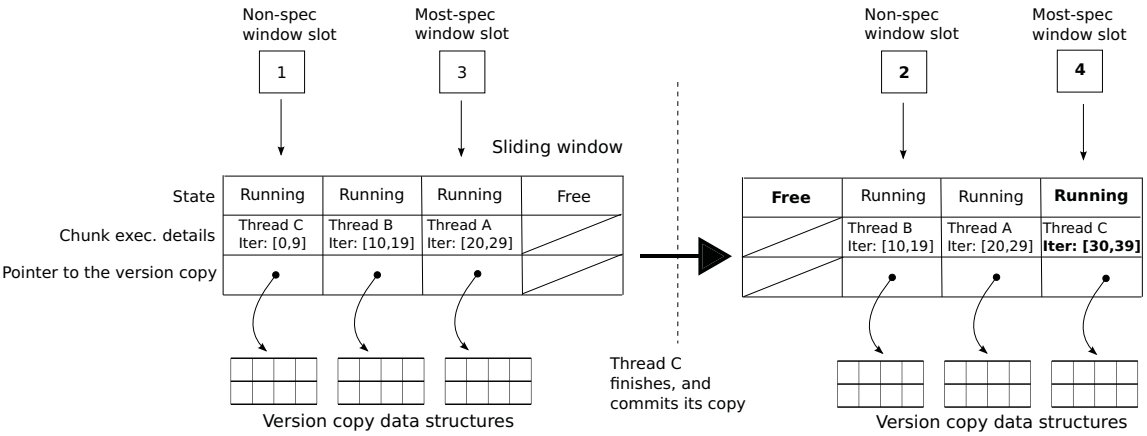


Figure 2: Updating the sliding window that handles the parallel, speculative execution. At a given moment (left), the thread C working in slot 1 is running. When Thread C finishes, it frees its slot and gets a new one, updating non-spec and most-spec pointers (right).

to each thread. ATLaS handles the parallel execution of each chunk of iterations through a sliding window mechanism, which is implemented by a matrix with W columns representing W window slots. Figure 2 depicts a simplified version of the sliding window implementation (see [4, 13] for more details). The figure represents a sliding window with four slots, hosting the execution of three parallel speculative threads.

The thread executing the earliest chunk of iterations (Thread C in our example) is called *non-speculative*, since it has no predecessors that may squash it. Conversely, the thread executing the latest chunk is called the *most-speculative* thread. As can be seen in Fig. 2, two pointers indicate the slots where the non-speculative and most-speculative threads are being executed. The part of the window being used is always the one from the non-spec pointer to the right, up to the most-spec pointer.

The only critical section in the ATLaS runtime library is the one that protects this sliding window. If two or more threads finish at the same time, they could be assigned to the same `Free` slot, resulting in an incorrect execution. Therefore, in order to ensure the correct operation of the ATLaS runtime library, it is necessary to protect the accesses to these shared structures, including the matrix that implements the sliding window mechanism, and the variables that point to the non- and most-speculative slots.

Figure 2 shows what happens when a non-speculative thread successfully finishes its execution. Suppose that Thread C, the one executing the non-speculative thread, finishes its execution and commits its data (the commit operation is not shown in the figure). After this, it enters the critical section to perform several actions. It marks slot 1 as Free; it advances the non-speculative pointer to slot 2; after checking that the slot past the most-speculative one is Free, it assigns it to itself, setting the most-speculative pointer to 4 and

changing its state to Running; and finally, after getting the following chunk of iterations to be executed (iterations 30 to 39 in our example), it exits the critical section. Note that the implementation of the sliding window works in a circular way: When Thread B eventually finishes, it will assign itself the slot that follows the one used by Thread C, in our case the leftmost slot.

The sliding window is modified in three different locations within the ATLaS runtime library. Therefore, the same lock is used in three different parts of the code to protect the access to these data structures. As will be seen, the place from where the access is performed has a noticeable impact in the performance of the protecting system being used. These places are the following:

- **(A) Each time a dependence violation is detected.** In the case of a write to a speculative variable, the thread in charge should update its version copy, and check whether a successor has consumed an outdated value of this variable. If this is the case, a *dependence violation* has happened, so the offending thread should be restarted in order to consume an updated version of the variable. This is done in several steps. First, the thread that has detected the situation should enter the critical section to change the state of the offending thread, from `Running` to `Squashed`, and the most-speculative pointer should be moved backwards to the last Running thread. After these changes, the thread exits the critical section and resumes its normal operation. The offending squashed thread will eventually discover its new state and will enter the critical section (see below).

- **(B) Each time a thread finishes its work,** either because the chunk has been successfully executed or because the thread discovers that it has been squashed. In both cases, the thread enters the critical section to change its own state from Running (resp. Squashed) to Free. After this operation, if the slot following the most-speculative one is Free, the thread assigns it to itself, and advances the most-speculative pointer by one. Otherwise, it means that the following slot is occupied either by a Running thread (this means that the window is full) or by another Squashed thread. In both cases our thread should exit the critical section and attempt to re-enter again, in order to give the thread that is using the slot the opportunity to free it[1] (see below).

- **(C) Each time a thread should wait for a free slot.** If a thread is not able to get a free window slot to work, because the following slot is not Free yet, it should get out and try to gain access again to the critical section to assign itself the following slot and to advance the most-speculative pointer.

[1] Our thread cannot simply wait inside the critical section, because it should get out in order to let the thread using that slot to get in and change its own state.

 ISBN: 978-84-608-6082-2

4 Protecting data accesses with OpenMP and STM

The original TLS runtime library uses the OpenMP critical directive to guarantee exclusive access of the threads to the three parts of the code mentioned above. Because the same data structures are accessed from three different places, the same lock is used to protect them in all cases. Recall that a block of code marked with an OpenMP critical directive is only executed by one thread at a time, whilst the rest of the threads that have reached the same point in the code have to wait. This procedure ensures that the sliding window is always in a consistent state, thus avoiding multiple threads concurrently updating this structure with the potential loss of consistency.

It is easy to see that the serialization of operations described above should imply a noticeable overhead in the performance of the speculative runtime library. A possible way to reduce this performance penalty would be to replace the strict, OpenMP critical construct with the more optimistic constructs that offer the Transactional Memory paradigm. The goal of STM is precisely to help in explicit parallel programming by reducing the costs of the locks required to avoid race conditions in critical sections [14, 15]. While OpenMP critical constructs only allow one single thread at a time inside the critical section, a transactional-based implementation allows several threads inside it, permitting their concurrent execution as long as consistency is not compromised.

However, the optimism of STM, as well as that of TLS, comes at the cost of some overheads, because of the extra instrumentation needed to handle the transactions, as well as the cost associated to the extra runs of particular transactions when a conflict appears. As can be seen, both OpenMP and STM approaches to protect data integrity have identified overheads. It is extremely difficult to predict which approach will be better for a particular problem, since it depends on the application, its running profile, and how often the benchmark accesses the potentially conflictive shared variables, among other factors.

Regarding the programmability, OpenMP has been designed to simplify, to a great extent, the process of parallelization, while the direct use of STM libraries involves a non-trivial instrumentation of the source code, from the definition of the transactional region to monitoring each access to speculative variables. This effort is mitigated by the existence of STM solutions that rely on the compiler to replace STM constructs with calls to the STM library. Some STM approaches propose language extensions or new constructs to declare transactional code regions that comprises statements that must be executed atomically. Then, either an ad-hoc compiler, or an existing compiler modified for this purpose, parses these new constructs, and generates all the instrumentation, in the same way as compilers process OpenMP constructs.

As we said above, OpenMP allows the user to delimit the critical sections with the construct `omp critical`. To declare a transactional region, STM libraries rely on different alternatives, such as new constructs (e.g. GCC-TM's `transaction_atomic{}` [16], the Intel's `tm_atomic{}` [17], or the more generic `transaction{}`), new compiler directives (such

Application	% target loop	Max. speedup P = 64 (Amhdahl)	% of iterations that present dep. violations	# of potentially speculative scalar variables	Size of chunks issued	Critical Sections accessed
FAST	100	64	0.001%	2	25	A, **B**
TREE	95.17	15.84	0%	259	100	**B**
2D-MEC	43.75	1.76	0.009%	10	1 800	A, **B**
2D-Hull, Kuzmin	100	64	0.0008%	1 206	11 000	A, **B**, C
2D-Hull, Square	100	64	0.0032%	3 906	3 000	A, **B**, C
2D-Hull, Disc	100	64	0.0219%	26 406	1 250	**A, B**, C
Delaunay	97.60	25.47	0.5%	12 030 060	2	**A, B, C**

Table 1: Percentages of potentially parallelism for the benchmarks and loops considered, together with some benchmarks' characteristics. Chunk sizes were selected to obtain maximum speedups.

as IBM's [18] `tm_atomic{}`), or even new OpenMP pragmas, such as `omp transaction`, defined by OpenTM [19]. Unfortunately, Intel STM compiler and OpenTM are not currently available, while the IBM compiler's transactional built-in memory functions are only valid for Power8 architecture and Blue Gene/Q.

In this work, we have used OpenMP, the GCC-TM, and the TinySTM libraries [20, 21] to protect the accesses to the sliding window described previously. These three approaches simplify the parallelization process with the mentioned constructs and directives. Moreover, GCC-TM defines a specification for transactional language constructs that other STM libraries can leverage, and hence, changing the underlying STM library is just a process of proper linking. In fact, TinySTM is compatible with GCC-TM, allowing programmers to use the same interface and save some programming effort.

Handling the critical sections with OpenMP is straightforward: The programmer should simply delimit the region by using the defined `omp critical` directive. This process is similar when using the GCC-TM specification. However, to ensure that the transaction is atomically executed, there may be certain functions inside the transaction that must *not* be executed. Since the compiler is not able to detect this issue for the functions called within a transaction, it is also necessary to annotate their declaration and specify whether they are safe to be called, with the `transaction_safe` attribute.

The following section describes the performance results obtained by ATLaS when using these three solutions to execute a set of real-world and synthetic benchmarks.

5 Experimentation

Experiments were carried out on a 64-processor server, equipped with four 16-core AMD Opteron 6376 processors at 2.3GHz and 256GB of RAM, which runs Ubuntu 12.04.3 LTS. All threads had exclusive access to the processors during the execution of the experiments, and we used wall-clock times in our measurements. We have used the OpenMP implementation

from GCC 4.8.2, and the transactional libraries from GCC-TM 4.8.2, and TinySTM 1.0.5.

To perform the experiments, we used both real-world and synthetic benchmarks. The real-world applications include the 2-dimensional Minimum Enclosing Circle (2D-MEC) problem [22], the 2-dimensional Convex Hull problem (2D-Hull) [23], the Delaunay Triangulation problem [24, 25], and a C implementation of the TREE benchmark [26]. We have also used a synthetic benchmark called Fast [4], which presents almost no dependences between iterations, and which was designed to test the overheads of the ATLaS runtime library.

Table 1 summarizes the characteristics of each benchmark, including the percentage of execution time consumed by each target loop, an estimation of the maximum speedup attainable (applying Amhdahls Law), the percentage of iterations of the target loop that lead to runtime dependence violations, the number of speculative variables within the loop, and the size of the chunk of consecutive iterations speculatively executed. I/O time consumed by the benchmarks were not taken into account. We also give an indication of which accesses to the sliding window protected by the critical section are more frequent in the benchmark (bold letters indicate that the corresponding call is more frequent). The performance results obtained by each benchmark and library used are summarized in Fig. 3.

The Fast benchmark was designed to test the efficiency of the speculative scheduling mechanism, with few iterations leading to a dependence violation, although they are enough to prevent a compiler from parallelizing the loop. This benchmark has very few dependence violations, so the critical section is primarily accessed to get the following chunk of iterations to be executed (access of type B in our library). As can be seen in the corresponding performance plot, OpenMP and the STM libraries handle the critical sections equally well, delivering almost identical performances, with a speedup of up to 37× with 64 processors.

Unlike the rest of the benchmarks, TREE does not suffer from dependence violations, but it is still not parallelizable at compile time because the compiler is not able to ensure that there are no data dependencies. Since it does not present dependence violations, the code that accesses the critical section is primarily B. Again, OpenMP and STM solutions deliver the same performance, with a peak speedup of 6× when running this benchmark with a 4096-point input set. As can be seen in the figure, the overheads of the TLS runtime library lead to a performance loss when using 48 threads or more, regardless of the implementation chosen to handle critical sections.

The 2D-MEC benchmark is a tricky code which has only 10 speculative variables that are frequently accessed. This benchmark calls the speculative loop many times with a very different number of iterations each time, making threads access the sliding window system frequently to get the following chunk. As long as it presents some dependence violations, the critical sections are accessed by codes A, but mostly B (C is rarely accessed in this benchmark). For this benchmark, the use of the OpenMP critical sections leads to the best performance, while the STM libraries leading to much poorer results. OpenMP gets a peak

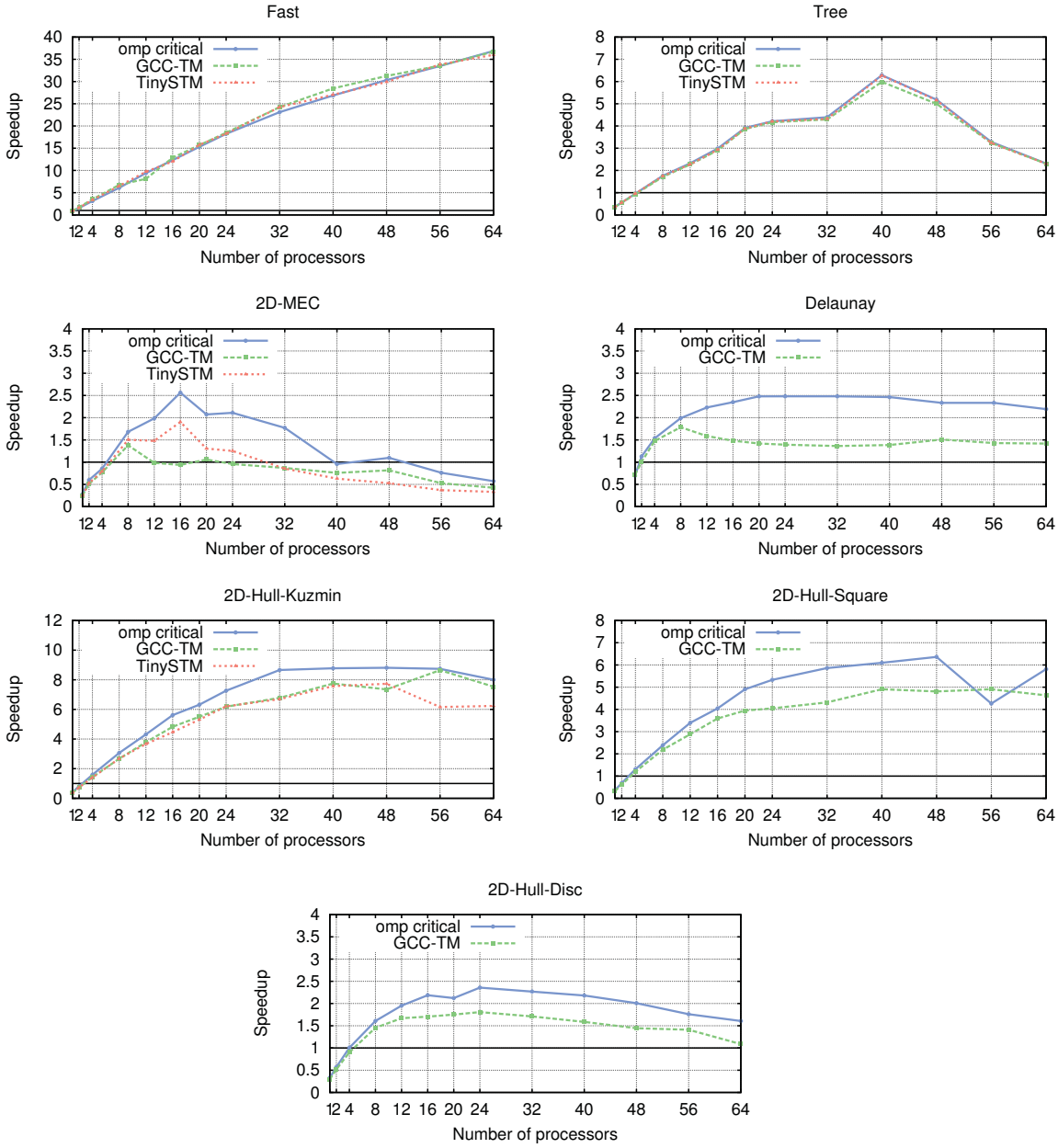


Figure 3: Speedups by number of processors for each tested benchmark, comparing the performance obtained by using OpenMP critical sections, and GCC-TM and TinySTM transactional libraries

speedup of 2.5×, outperforming the speedup of TinySTM (1.9×) and GCC-TM (1.3×).

The overheads of TinySTM and GCC-TM are even more noticeable for the Delaunay problem, the benchmark with the highest number of dependence violations (0.5%) and the smallest optimum chunk size (just two iterations). Therefore, critical sections are accessed quite frequently, because of the detection of dependence violations (code A), the need to schedule the execution of many chunks (code B), and some degree of load imbalance which frequently makes the window to be full, leading to contention (code C). The stress on the exclusive access management is so high that, in fact, the TinySTM library is not able to properly handle the accesses to the protected sliding window when running this problem with two or more threads. Hence, we cannot present performance figures for this library in this case. Meanwhile, GCC-TM leads to very modest speedups compared with the use of OpenMP critical sections.

A similar issue occurs with the execution of the 2D-Hull problem with different input sets. We have found that runs of the 2D-Hull problem, with datasets whose execution involves a larger amount of conflicts and dependence violations (as happen when the Square and Disc input sets are used), do not finish when using TinySTM, regardless of the number of parallel threads. TinySTM also fails when using 2 threads and the Kuzmin dataset, producing different, unexpected outcomes on each execution. Besides this, although both STM libraries perform similarly for each dataset, their performance is consistently worse than the one offered by the OpenMP critical sections.

From the results described above we can make the following observations:

1. Not surprisingly, the relative performance of the lock implementations depends, to a great extent, on the running profile and the particular characteristics of each benchmark. In our case, this profile not only depends on the number of dependence violations (that requires accesses through code A to squash the offending threads), but the number of chunks of iterations that should be scheduled (requiring accesses to codes B and C).

2. In general terms, the more frequent the accesses to the protected data structures, the poorer the performance of the STM libraries with respect to the OpenMP critical sections. In fact, if the number of accesses is high enough, TinySTM starts to fail. In this sense, we have found that the GCC-TM implementation is more robust.

3. While we expected that the performance of the STM implementations would decay when the number of accesses to the protected structures increases, we also expected that, when the number of accesses were relatively low, the performance obtained with STM would be better than using the OpenMP critical sections. As our performance figures show, this is not the case, the use of the OpenMP critical sections being the best choice for our problem.

6 Conclusions

The aims of this study were to test whether the use of STM might lead to an improvement in the performance of our software-based, thread-level speculation system, and to assess the efficiency and maturity of STM libraries. Our experimental results show that, in general, STM solutions deliver poorer performance than the use of the classical OpenMP critical sections for our problem. Regarding the maturity of STM libraries, in our experience, the GCC-TM implementation is mature enough to be used for production purposes, while TinySTM still presents room for improvement. Our present and future work includes testing the performance of hardware TM solutions in this same context.

7 Acknowledgments

This research has been partially supported by MICINN (Spain) and ERDF program of the European Union: HomProg-HetSys project (TIN2014-58876-P), CAPAP-H5 network (TIN2014-53522-REDT), and COST Program Action IC1305: Network for Sustainable Ultrascale Computing (NESUS).

References

[1] Chandra, R., Menon, R., Dagum, L., Kohr, D., Maydan, D., McDonald, J.: Parallel Programming in OpenMP. 1 edn. Morgan Kaufmann (October 2000)

[2] Shavit, N., Touitou, D.: Software transactional memory. Distributed Computing **10** (1997) 99–116

[3] Wong, M., Bihari, B.L., de Supinski, B.R., Wu, P., Michael, M., Liu, Y., Chen, W.: A case for including transactions in OpenMP. IWOMP'10 Proceedings (2010) 149160

[4] Estebanez, A., Aldea, S., Llanos, D.R., Gonzalez-Escribano, A.: An OpenMP extension that supports thread-level speculation. IEEE Transactions on Parallel and Distributed Systems, to appear.

[5] Gupta, M., Nim, R.: Techniques for speculative run-time parallelization of loops. In: Proc. of the 1998 ACM/IEEE Conference on Supercomputing. (1998) 1–12

[6] Rauchwerger, L., Padua, D.: The LRPD test: Speculative run-time parallelization of loops with privatization and reduction parallelization. In: PLDI'95 Proceedings. (1995) 218–232

[7] Cintra, M., Llanos, D.R.: Toward efficient and robust software speculative parallelization on multiprocessors. In: PPoPP'03 Proceedings. (June 2003) 13–24

[8] Dang, F.H., Yu, H., Rauchwerger, L.: The R-LRPD test: Speculative parallelization of partially parallel loops. In: 16th IPDPS Proceedings. (2002) 20–29

[9] Xekalakis, P., Ioannou, N., Cintra, M.: Combining thread level speculation helper threads and runahead execution. In: ICS 2009 Proceedings. (2009) 410–420

[10] Gao, L., Li, L., Xue, J., Yew, P.C.: SEED: A statically greedy and dynamically adaptive approach for speculative loop execution. IEEE Transactions on Computers **62**(5) (2013) 1004–1016

[11] Kulkarni, M., Pingali, K., Walter, B., Ramanarayanan, G., Bala, K., Chew, L.P.: Optimistic parallelism requires abstractions. In: PLDI Proceedings. (2007) 211–222

[12] Aldea, S., Estebanez, A., Llanos, D.R., Gonzalez-Escribano, A.: A new gcc plugin-based compiler pass to add support for thread-level speculation into openmp. In: Euro-Par 2014 Proceedings. LNCS 8632 (2014) 234–245

[13] Estebanez, A., Llanos, D.R., Gonzalez-Escribano, A.: New data structures to handle speculative parallelization at runtime. In: HLPP '14 Proceedings. (2014)

[14] Ceze, L., Tuck, J., Torrellas, J., Cascaval, C.: Bulk disambiguation of speculative threads in multiprocessors. ACM SIGARCH Computer Architecture News **34**(2) (2006) 227–238

[15] Barreto, J., Dragojevic, A., Ferreira, P., Filipe, R., Guerraoui, R.: Unifying thread-level speculation and transactional memory. In: Middleware '12 Proceedings. (2012) 187–207

[16] Riegel, T.: Transactional Memory in GCC. `https://gcc.gnu.org/wiki/TransactionalMemory` (2012)

[17] : Intel C++ STM Compiler, Prototype Edition (2012)

[18] IBM: Thread-level speculative execution for C/C++ (2012) Tech. report.

[19] Baek, W., Minh, C.C., Trautmann, M., Kozyrakis, C., Olukotun, K.: The OpenTM transactional application programming interface. In: 16th ISCA Proceedings, IEEE Computer Society (2007) 376–387

[20] Pascal Felber, Christof Fetzer, P.M., Riegel, T.: Time-based software transactional memory. IEEE Transactions on Parallel and Distributed Systems **21**(12) (December 2010) 1793–1807

[21] Pascal Felber, C.F., Riegel, T.: Dynamic performance tuning of word-based software transactional memory. In: PPoPP '08 Proceedings . (2008) 237–246

[22] Welzl, E.: Smallest enclosing disks (balls and ellipsoids). In: New results and new trends in computer science. Volume 555 of Lecture Notes in Computer Science., Springer-Verlag (1991) 359–370

[23] Clarkson, K.L., Mehlhorn, K., Seidel, R.: Four results on randomized incremental constructions. Comput. Geom. Theory Appl. **3**(4) (1993) 185–212

[24] Mücke, E.P., Saias, I., Zhu, B.: Fast randomized point location without preprocessing in two- and three-dimensional Delaunay triangulations. In: SoCG '96 Proceedings. (1996) 274–283

[25] Devroye, L., Mücke, E.P., Zhu, B.: A note on point location in Delaunay triangulations of random points. Algorithmica **22** (1998) 477–482

[26] Barnes, J.E.: TREE. Institute for Astronomy. University of Hawaii. ftp://hubble.ifa.hawaii.edu/pub/barnes/treecode/ (1997)

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Fixed Point Theorems for Graph Dynamical Systems

Juan A. Aledo[1], Luis G. Díaz[1], Silvia M. Sanahuja[1] and Jose C. Valverde[1]

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

emails: juanangel.aledo@uclm.es, LuisGabriel.Diaz@alu.uclm.es, Silvia.MSanahuja@uclm.es, Jose.Valverde@uclm.es

Abstract

In this work, we provide conditions to obtain fixed point theorems for both parallel and sequential graph dynamical systems with (Boolean) maxterms and minterms as global evolution operators. In order to do that, we previously study the orbital structure of such discrete dynamical systems.

Key words: parallel dynamical system, sequential dynamical system, orbital structure, fixed point theorem

1 Introduction

In the last two decades, *graph dynamical systems* (GDS) have revealed as an important tool for the mathematical modeling of computer processes. For the formulation of such models, these systems are decomposed in the lower unities of aggregation, called *entities*, where each entity i has a (numerical) value x_i which represents its state. When the possible states of each entity are activated or deactivated, the state value of an entity i is formalized by considering $x_i \in \{0, 1\}$. The relations among entities are represented by a graph called the *dependency graph* of the system, while the evolution or update of the system is implemented by *local (Boolean) functions* which together constitute a global (evolution) operator. That is, for the dynamic evolution of the state of any entity, the corresponding local function acts on such state and those ones corresponding to entities related to it.

When the states of the entities are updated in a synchronous manner, the system is called a *parallel dynamical system* (PDS) [2, 5], while if they are updated in an asynchronous way, the system is named *sequential dynamical system* (SDS) [6].

The main aim regarding the study of a dynamical system is to give a complete characterization of its orbit structure. That is, to derive as much information as possible about

 ISBN: 978-84-608-6082-2

the phase diagram, based on the initial states and the evolution operator of the system. Specifically, it means to determine the length and number of coexisting limit cycles and which different initial states arrive in the same limit cycle.

In this sense, for the case of PDS with a maxterm or minterm as evolution operator it was shown in [2] that the periodic orbits of such graph dynamical systems are fixed points or 2-periodic orbits.

The main objective of this paper is to establish results for PDS and SDS with general maxterm or minterm as evolution operators regarding their orbital structure, and in particular in the sense of the Fixed Point Theorem by Banach.

2 Results

Next, we briefly enumerate the results that make up our talk.

- For PDS [3]:
  - We describe which structure these systems must have in order to admit fixed points.
  - We show that fixed points and (eventually) 2-periodic orbits cannot coexist.
  - We find conditions to assure that these systems have a unique fixed point (Fixed Point Theorem).
- Regarding SDS [4]:
  - We prove that, in contrast with the case of PDS, periodic orbits of any period can appear.
  - We demonstrate that the coexistence of orbits of any different periods greater than or equal to two is possible, although we also prove that the existence of fixed points excludes the presence of other periodic orbits.
  - We give a characterization of SDS for which the unique periodic orbits are fixed points.
  - We provide conditions that characterize the uniqueness of a fixed point and establish a Fixed Point Theorem.

Acknowledgements

Juan A. Aledo is supported by Junta de Comunidades de Castilla-La Mancha Grant PEII-2014-049-P. Silvia M. Sanahuja and Jose C. Valverde are supported by the Ministry of Economy and Competitiveness of Spain Grant MTM2014-51891-P and the University of Castilla-La Mancha Grant GI20152987.

References

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

[2] J.A. Aledo, S. Martínez, F.L. Pelayo and J.C. Valverde, *Parallel dynamical systems on maxterm and minterm Boolean functions*, Math. Comput. Model. **35** (2012) 666–671.

[3] J.A. Aledo, L.G. Díaz, S.M. Sanahuja and J.C. Valverde, *Fixed Point Theorems for Parallel Dynamical Systems over Graphs*, Preprint.

[4] J.A. Aledo, L.G. Díaz, S.M. Sanahuja and J.C. Valverde, *Orbital Structure and Fixed Point Theorem for Sequential Dynamical Systems*, Preprint.

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

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

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A fractional Malthusian growth model with variable order using an optimization approach

Ricardo Almeida[1], Nuno R. O. Bastos[1,2] and M. Teresa T. Monteiro[3]

[1] *Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810–193 Aveiro, Portugal*

[2] *Department of Mathematics, School of Technology and Management of Viseu, Polytechnic Institute of Viseu, 3504–510 Viseu, Portugal*

[3] *Algoritmi R&D Center, Department of Production and Systems, University of Minho, Campus de Gualtar, 4710–057 Braga, Portugal*

emails: ricardo.almeida@ua.pt, nbastos@estv.ipv.pt, tm@dps.uminho.pt

Abstract

The goal of this work is to show, based on concrete data, that fractional differential equations with variable fractional order are more efficient to model the world population growth than the classical differential equation, or even a fractional differential equation with constant order. With these new models, we can predict more efficiently the population growth based on the present data.

Key words: Fractional calculus, fractional differential equation, least squares, unconstrained optimization

1 Introduction

The Malthusian growth model was proposed in 1798 by the English economist Thomas Malthus in his book *An Essay on the Principle of Population*. The theory states that the population number has exponential growth based on a constant rate, applied to ideal circumstances or to a short period of time, when an individual lives in region with no constraints on food and with no natural enemies. In this case, if $N(t)$ represents the size of the population at an instant t, the dynamic differential equation

$$N'(t) = P \cdot N(t)$$

models the growth of the population. The constant P, called the Malthusian parameter, is given by the difference between the fertility and the mortality rates, assuming that these rates are constant in time. If N_0 is the initial level of the population, the function

$$N(t) = N_0 \exp(Pt) \tag{1}$$

gives the exact number of individuals at a given time t. Although there are several models to describe the dynamics of the population growth, the Malthusian model has the advantage that it is given by a linear differential equation. Later, when we model the same problem but using a fractional differential equation, we know the analytic expression of solution to the problem.

To test the different models that we purpose here, we will see how close they are to real data, by fitting the solution with dependence on some parameters with the observations. One of the most used methods is the least squares technique. Suppose that the data consists in m points, say $(t_1, x_1), \ldots, (t_m, x_m)$, and we intend to fit these values in a theoretical model $t \mapsto x(t)$, where the form x is known but it depends on some unknown parameters $\beta_1, \ldots, \beta_k$. If we consider in each step the error $d_i := x_i - x(t_i)$, for $i = 1, \ldots, m$, then the total error is given by

$$E := \sum_{i=1}^{m} (d_i)^2.$$

The goal is to find the values of the parameters $\beta_1, \ldots, \beta_k$ for which E attains a minimum value.

2 World population

In [1], a fractional approach was considered to model the World Population Growth. Starting with the classical model

$$N'(t) = P \cdot N(t), \tag{2}$$

the ordinary derivative was replaced by the Caputo fractional derivative, and the dynamic was described by the fractional differential equation (for fractional calculus theory, see [2, 3])

$${}^{C}D^{\alpha}_{0+} N(t) = P \cdot N(t), \quad t \geq 0,\ \alpha \in (0, 1). \tag{3}$$

The solution to the fractional problem is given by the function

$$N(t) = N_0 E_{\alpha}(Pt^{\alpha}), \tag{4}$$

where $E_{\alpha}(\cdot)$ denotes the Mittag–Leffler function:

$$E_{\alpha}(z) := \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(k\alpha + 1)}, \quad z \in \mathbb{R}.$$

Then, using the data available from the United Nations [4] from year 1910 until 2010, the best values for the parameters were found. The objective is to minimize the sum of the squares of the offsets, and to get a better accuracy for the model, the fractional order α was considered free, without any constraints. For the classical model (2), the values obtained were

$$P \approx 1.3501 \times 10^{-2} \quad \text{with the error} \quad E_{classical} \approx 7.0795 \times 10^{5}.$$

When we considered the problem modeled by the fractional differential equation (3), the values were

$$\alpha \approx 1.3933, \quad P \approx 3.4399 \times 10^{-3} \quad \text{with the error} \quad E_{fractional} \approx 2.0506 \times 10^{5}.$$

So, from these results, we see that the fractional approach is more efficient in modelling the problem than the ordinary one. The next step is to consider even a more general approach to this problem, by considering the fractional order to be a function depending on time $t \mapsto \alpha(t)$. Motivated by Eq. (4), and considering the Mittag–Leffler function with variable order

$$E_{\alpha(t)}(z) := \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(k\alpha(t)+1)}, \quad z \in \mathbb{R},$$

we purpose the following theoretical model to study the world population problem:

$$N(t) = N_0 E_{\alpha(t)}(Pt^{\alpha(t)}). \tag{5}$$

Observe that, when $\alpha(\cdot)$ is constant, $\alpha(t) \equiv \alpha$, then Eq. (5) reduces to Eq. (4), which in turn when $\alpha \to 1$, we obtain the classical model (1). We test model (5) by the closeness to the observed data, from which we infer the values of the parameters. We compare the fractional model with constant order with a new one, with variable fractional orders. For example, when considering the order

$$\alpha_1(t) := at^2 + bt + c$$

the best values for this fractional order are

$$a \approx -4.4865 \times 10^{-5}, \quad b \approx 7.5332 \times 10^{-3}, \quad c \approx 8.5596 \times 10^{-1} \quad \text{and} \quad P \approx 7.5849 \times 10^{-3}$$

with error $E \approx 1.4813 \times 10^{4}$. The results are shown in Figure 1.

Acknowledgements

The first and second authors were supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), within project UID/MAT/04106/2013; third author by the ALGORITMI R&D Center and project COMPETE: POCI-01-0145-FEDER-007043 and FCT Fundação para a Ciência e a Tecnologia within the Project Scope: UID/CEC/00319/2013.

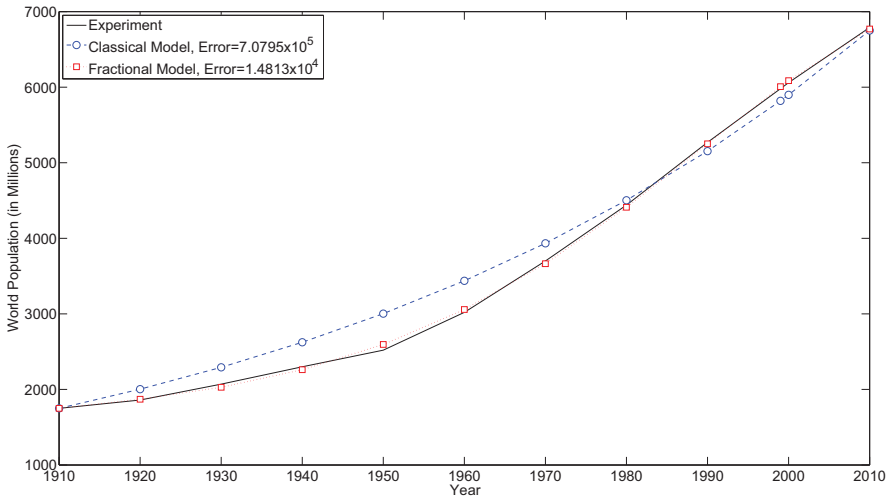


Figure 1: World Population: data, classical and fractional models.

References

[1] R. Almeida, N.R.O. Bastos and M.T.T. Monteiro. Modelling some real phenomena by fractional differential equations. Math. Meth. Appl. Sci. (in press)

[2] A.B. Malinowska, T. Odzijewicz and D.F.M. Torres, Advanced Methods in the Fractional Calculus of Variations, Springer Briefs in Applied Sciences and Technology, Springer, Cham, 2015.

[3] S. G. Samko, A. A. Kilbas and O. I. Marichev, Fractional integrals and derivatives, translated from the 1987 Russian original, Gordon and Breach, Yverdon, 1993.

[4] United Nations. The World at Six Billion Off Site, Table 1, World Population From Year 0 to Stabilization, 5, 1999.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Backward error analysis of almost strictly sign regular matrices

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

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

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

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

Abstract

Sign regular matrices are matrices whose minors of the same order have the same sign. A sign regular matrix is almost strictly sign regular if all its nontrivial minors of the same order have the same strict sign. In this paper, componentwise backward error analysis for this kind of matrices is performed when Neville elimination with two-determinant pivoting strategy is applied.

Key words: sign regular matrices, almost strictly sign regular matrices, backward error

MSC 2000: 65F05, 65F15, 65F35

1 Introduction

Error analysis of Gauss elimination is a subject that became firmly established fifty years ago due specially to Wilkinson's work (see [14] and [15]). Let us assume that $Ax = b$ is a linear system where $A = LU$ (L and U are lower and upper triangular matrices respectively) is a nonsingular matrix and that Gauss elimination applied to A in floating point arithmetic has produced a factorization $\hat{L}\hat{U}$ and a solution $\hat{x}$.

To investigate the effect of rounding errors when working with floating point arithmetic we use the classic model

$$\begin{aligned} fl(x\ op\ y) &= (x\ op\ y)(1+\delta), & |\delta| \leq u, \\ fl(x\ op\ y) &= \frac{(x\ op\ y)}{(1+\epsilon)}, & |\epsilon| \leq u, \end{aligned} \tag{1}$$

 ISBN: 978-84-608-6082-2

where u is the unit roundoff and op denotes any of the operations $\{+, -, *, /\}$. In addition, we assume that neither overflow nor underflow occurs.

Wilkinson's componentwise backward error analysis of Gauss elimination without pivoting was stated (see [14]) in the following way:

$$\hat{L}\hat{U} = A + E, \quad |E| \leq \gamma_n |\hat{L}||\hat{U}|, \tag{2}$$

where $\gamma_n := \dfrac{nu}{1-nu}$ and u is the unit roundoff.

Note that, in certain cases $|\hat{L}||\hat{U}| = |\hat{L}\hat{U}|$, which arises, for instance, when both $\hat{L}$ and $\hat{U}$ are nonnegative. In this situation

$$|\hat{L}\hat{U}| = |\hat{L}||\hat{U}| = |A + E| \leq |A| + \gamma_n |\hat{L}||\hat{U}|,$$

and thereby

$$|\hat{L}||\hat{U}| \leq \frac{1}{1-\gamma_n}|A|, \tag{3}$$

(see section 9.2 of [9]).

One class of matrices that has nonnegative LU factors is defined as follows. Totally positive (TP) matrices are matrices whose minors are nonnegative. These matrices arise naturally en many areas of mathematics, economics, etc. Their applications to approximation theory and computer geometric design are specially interesting (see [4], [7], [12] and [13]).

If the matrix A is nonsingular TP (see [5]), de Boor and Pinkus showed that the componentwise relative backward error is pleasantly small. They obtained, in that case, that the backward error matrix E can be chosen to satisfy the inequality

$$|E| \leq \frac{\gamma_n}{1-\gamma_n} A, \tag{4}$$

which follows from (2) and (3).

If we consider now the equation $(A+H)\hat{x} = b$, the corresponding formulas derived from (2) and (3) (see Theorem 9.4 of [9]) are

$$|H| \leq 2\gamma_n |\hat{L}||\hat{U}|, \tag{5}$$

$$|H| \leq \frac{2\gamma_n}{1-\gamma_n} A. \tag{6}$$

In some papers (see [2] and [8]) it has been shown that an elimination procedure, called Neville elimination (NE), is very convenient when working with TP matrices and other related type of matrices. If A is a nonsingular $n \times n$ matrix, NE consists of at most $n-1$ successive major steps, resulting in a sequence of matrices as follows:

$$A = \widetilde{A}^{(1)} \to A^{(1)} \to \cdots \to \widetilde{A}^{(n)} = A^{(n)} = U \tag{7}$$

where U is an upper triangular matrix.

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

$$a_{it}^{(t)} = 0,\ i \geq t \Rightarrow a_{ht}^{(t)} = 0,\ \forall h \geq i. \tag{8}$$

The matrix $A^{(t)}$ is obtained from $\widetilde{A}^{(t)}$ reordering rows t, $t+1$, ..., n according to a row pivoting strategy that satisfies (8). To obtain $\widetilde{A}^{(t+1)}$ from $A^{(t)}$ we produce zeros in the column t below the main diagonal by subtracting a multiple of the ith row from the $(i+1)$th, for $i = n-1, n-2, \ldots, t$, according to the following formula:

$$\widetilde{a}_{ij}^{(t+1)} = \begin{cases} a_{ij}^{(t)}, & 1 \leq i \leq t, \\ a_{ij}^{(t)} - \frac{a_{it}^{(t)}}{a_{i-1,t}^{(t)}} a_{i-1,j}^{(t)}, & \text{if } a_{i-1,t}^{(t)} \neq 0, \quad t+1 \leq i \leq n, \\ a_{ij}^{(t)}, & \text{if } a_{i-1,t}^{(t)} = 0, \quad t+1 \leq i \leq n, \end{cases} \tag{9}$$

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

The element

$$p_{ij} = a_{ij}^{(j)}, \quad 1 \leq j \leq i \leq n, \tag{10}$$

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

$$m_{ij} = \begin{cases} \dfrac{a_{ij}^{(j)}}{a_{i-1,j}^{(j)}} \left(= \dfrac{p_{ij}}{p_{i-1,j}} \right), & \text{if } a_{i-1,j}^{(j)} \neq 0, \\ 0, & \text{if } a_{i-1,j}^{(j)} = 0, \end{cases} \tag{11}$$

the (i,j) multiplier. When A is a nonsingular TP matrix, then NE can be performed without row exchanges (see [8]).

In [1], the authors study the backward error analysis of NE when the NE can be performed without row exchanges, for both the general case and for the case where the coefficient matrix of the linear system is TP. In this last situation (see Theorem 4.1 of [1]), they showed that

$$|E| \leq \beta_n \hat{L}\hat{U} \leq \frac{\beta_n}{1-\beta_n} A, \tag{12}$$

$$|E| \leq \alpha_n A, \tag{13}$$

$$|E| \leq \gamma_{n-1} A, \tag{14}$$

with $\gamma_n := \dfrac{nu}{1-nu}$, $\beta_n := (1+u)^{n-1} - 1$ and $\alpha_n := (1-u)^{1-n} - 1$. Note that $\alpha_n \leq \gamma_{n-1}$.

More recently, Huang et *al.* (see [11]) have performed a componentwise backward error analysis of NE whenever there need row exchanges.

Let A be a nonsingular real matrix $n \times n$. Taking into account (7), the matrix A can be factorized as

$$A = Q_1 L_1 Q_2 L_2 \cdots Q_{n-1} L_{n-1} U,$$

where U is upper triangular, and for all $t = 1, 2, \ldots, n-1$, Q_t is the permutation matrix corresponding to (8) and

$$L_t = E_n(m_{nt}) E_{n-1}(m_{n-1,t}) \ldots E_{t+1}(m_{t+1,t}).$$

As in [8], we denote by $E_i(\alpha)$ the bidiagonal lower triangular matrix whose (r, s) entry $(1 \leq r, s \leq n)$ is given by

$$\begin{cases} 1, & \text{if } r = s, \\ \alpha, & \text{if } (r, s) = (i, i-1), \\ 0, & \text{elsewhere.} \end{cases} \tag{15}$$

In particular (see Theorem 7 of [11]), when they consider A nonsingular sign regular (SR) matrix (matrices whose minors of the same order have the same sign) and apply NE with two-determinant pivoting to A in finite float point arithmetic, under condition (24) of [11], the obtained result is

$$|E| \leq \frac{\Psi_n}{1 - \Psi_n} |A|, \tag{16}$$

where $\Psi_n = (1+u)^{\frac{n(n-1)}{2}} - 1$.

Almost strictly sign regular (ASSR) matrices are SR matrices whose nontrivial minors of the same order have all the same strict sign (see following section and [10]). In this work, some interesting properties related with the application of NE with two-determinant pivoting strategy to ASSR matrices are presented. Besides, componentwise backward error analysis for this type of matrices is performed.

2 On the computed matrices

In this section we analyze some aspects with regard to the application of the NE with two-determinant pivoting strategy to ASSR matrices.

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

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

So, A is a type-II staircase matrix if it satisfies that P_nA is a type-I staircase matrix, where P_n is the backward identity matrix $n \times n$.

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

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

Next, ASSR matrices are defined:

Definition 2 *A real $n \times n$ matrix A is said to be ASSR with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ if t is either type-I or type-II staircase and all its nontrivial minors* $\det A[\alpha|\beta]$ *satisfy that*

$$\varepsilon_m \det A[\alpha|\beta] > 0, \quad \alpha, \beta \in Q_{m,n}, \qquad m \leq n. \tag{17}$$

Observe that an ASSR matrix is nonsingular.

In [6] a row pivoting strategy associated to NE for nonsingular SR matrices is introduced. It has be called two-determinant pivoting strategy.

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

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

The following result corresponds to Proposition 1 of [3].

Proposition 1 *Let $A = (a_{1\leq i,j\leq n})$ be an ASSR type-I staircase matrix, with zero pattern $J = \{j_0, j_1, \ldots, j_{l-1}, j_l\}$ and $l \geq 2$. Then, the NE with two-determinant pivoting strategy does not involve row exchanges until the step $t = j_{l-1}$.*

 ISBN: 978-84-608-6082-2

The next result (see Theorem 2 of [3]) shows that almost strict sign regularity is inherited by all matrices $\widetilde{A}^{(t)}[t,\ldots,n]$ when we apply NE with two-determinant pivoting.

Theorem 1 *Let $A=(a_{ij})_{1\leq i,j,\leq n}$ be an ASSR matrix, and let us apply NE with two-determinant pivoting strategy. Then, for all $t\in\{1,\ldots,n\}$, all matrices $\widetilde{A}^{(t)}[t,\ldots,n]$ are ASSR and $\varepsilon_1(A)=\varepsilon_1(\widetilde{A}^{(t)})$.*

Remark 1 *Taking into account the previous result, we can show that $A^{(1)}[2,\ldots,n]$ and $\widetilde{A}^{(2)}[2,\ldots,n]$ have the zero entries exactly in the same positions. Since the first row is unchanged, we can state that $\widetilde{A}^{(2)}$ has zeros placed in the same positions as $A^{(1)}$, except those arising from the process of NE.*

Theorem 1 allows us to show some properties about $\varepsilon_1(U)$ and the multipliers of NE.

Remark 2 *Note that if A is an ASSR matrix, and we apply NE with two-determinant pivoting strategy, by Theorem 1 all matrices $\widetilde{A}^{(t)}[t,\ldots,n]$ are ASSR and $\varepsilon_1(A)=\varepsilon_1(\widetilde{A}^{(t)})$. So,*

$$\varepsilon_1(U)=\varepsilon_1(A),$$

and for all $t=1,2,\ldots,n-1$,

$$m_{nt}\geq 0,\ m_{n-1,t}\geq 0,\ m_{t+1,t}\geq 0.$$

Next, we consider the case where A is an ASSR matrix and we perform a backward error analysis of NE with two-determinant pivoting strategy, assuming sufficiently high finite precision. For this purpose, some auxiliary results are presented.

Finite precision arithmetic produces a sequence of matrices $\widehat{\widetilde{A}^{(t)}}$ in the NE of A that can be different from the sequence $\widetilde{A}^{(t)}$ obtained by exact arithmetic. Nevertheless, taking into account (9) and (1) we assume the fact (provable by induction on n) that for sufficiently small unit roundoff u

$$\widehat{\widetilde{A}^{(t)}}\to\widetilde{A}^{(t)}\quad\text{as}\quad u\to 0. \tag{18}$$

Next, we present some results about nonzero entries of ASSR matrices when the NE with two-determinant pivoting strategy is performed.

Theorem 2 *Let $A=(a_{ij})_{1\leq i,j\leq n}$ be an ASSR matrix with signature $\varepsilon=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)$ with $\varepsilon_2=+1$. If NE with two-determinant pivoting strategy is applied to A, then*

$$a_{ij}\neq 0,\ i\leq j\ \Rightarrow\ \widetilde{a}_{ij}^{(t)}\neq 0,\ t=1,\ldots,n, \tag{19}$$

and

$$a_{ij}\neq 0,\ i>j\ \Rightarrow\ \widetilde{a}_{ij}^{(t)}\neq 0,\ t=1,\ldots,j. \tag{20}$$

Corollary 1 *Let $A = (a_{ij})_{1\leq i,j\leq n}$ be an ASSR matrix. If NE with two-determinant pivoting strategy is applied to A, then*

$$a_{ij}^{(1)} \neq 0,\ i \leq j \ \Rightarrow\ \widetilde{a}_{ij}^{(t)} \neq 0,\ t = 1, \ldots, n, \tag{21}$$

and

$$a_{ij}^{(1)} \neq 0,\ i > j \ \Rightarrow\ \widetilde{a}_{ij}^{(t)} \neq 0,\ t = 1, \ldots, j. \tag{22}$$

Remark 3 *Notice that, under the assumptions of Theorem 1, we have that $\varepsilon_1(A) = \varepsilon_1(\widetilde{A}^{(t)})$. Besides, by Corollary 1, if $\varepsilon_1(A) = 1$, then*

$$a_{ij}^{(1)} > 0,\ i \leq j \ \Rightarrow\ \widetilde{a}_{ij}^{(t)} > 0,\ t = 1, \ldots, n, \tag{23}$$

and

$$a_{ij}^{(1)} > 0,\ i > j \ \Rightarrow\ \widetilde{a}_{ij}^{(t)} > 0,\ t = 1, \ldots, j. \tag{24}$$

Otherwise (if $\varepsilon_1(A) = -1$),

$$a_{ij}^{(1)} < 0,\ i \leq j \ \Rightarrow\ \widetilde{a}_{ij}^{(t)} < 0,\ t = 1, \ldots, n, \tag{25}$$

and

$$a_{ij}^{(1)} < 0,\ i > j \ \Rightarrow\ \widetilde{a}_{ij}^{(t)} < 0,\ t = 1, \ldots, j. \tag{26}$$

Taking into account the previous results and observations, the following result can be deduced:

Theorem 3 *Let A be an ASSR matrix with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$. Then, the computed matrices $(\widehat{\widetilde{A}^{(t)}})$, for a sufficiently small unit roundoff, satisfy*

$$\varepsilon_1\left(\widehat{\widetilde{A}^{(t)}}\right) = \varepsilon_1(A), \quad t = 1, \ldots, n. \tag{27}$$

3 Backward error analysis of NE with two-determinant pivoting

Taking into account the previous results, we perform a backward error analysis of NE with two-determinant pivoting strategy when A is an ASSR matrix.

By Theorem 3, the hypotheses of Theorem 7 of [11] are satisfied for ASSR matrices (which form a subclass of SR matrices) for a sufficiently small unit roundoff. Therefore (16) is satisfied. Now, we shall improved the bound (16) for ASSR matrices. If A is an ASSR matrix type-I staircase matrix, with zero pattern $J = \{j_0, j_1, \ldots, j_{l-1}, j_l\}$ and $l \geq 2$,

 ISBN: 978-84-608-6082-2

by Proposition 1, the NE with two-determinant pivoting strategy does not involve rows exchanges until the step $t = j_{l-1}$. Then

$$Q_1 = Q_2 = \cdots = Q_{j_{l-1}} = I,$$

and the bound (16) becomes:

$$\widehat{F}_1\widehat{F}_2 \ldots \widehat{F}_{j_{l-1}} Q_{j_l} \widehat{L}_{j_l} \ldots Q_{n-1}\widehat{L}_{n-1}\widehat{U} = A + E, \tag{28}$$

where it can be proved that

$$|E| \leq C(n, u, j_{l-1})[|\widehat{F}_1| \ldots |\widehat{F}_{j_{l-1}}| Q_{j_l} |\widehat{L}_{j_l}| \ldots Q_{n-1}|\widehat{L}_{n-1}|\widehat{U}|] \tag{29}$$

where $C(n, u, j_{l-1})$ is a function that depends of the order of A, the unit roundoff and the index j_{l-1}.

In addition,

$$|E| \leq \frac{C(n, u, j_{l-1})}{1 - C(n, u, j_{l-1})}|A|. \tag{30}$$

Note that the matrices $F_i := E_{n-(i-1)}(m_{n-(i-1),1}) \cdots E_n(m_{ni})$, with $i = 1, \ldots, j_{l-1}$, are obtained as

$$F_i = \begin{pmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ & \ddots & \ddots & & & & & & \\ & & 0 & 1 & & & & & \\ & & & m_{n-(i-1),1} & 1 & & & & \\ & & & & m_{n-(i-2),2} & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & m_{ni} & 1 \end{pmatrix}. \tag{31}$$

Now, we assume that A is an ASSR matrix satisfying the WR condition (NE can be performed without changes of rows) and that we compute the NE of A with sufficiently high finite precision so that it can be carried out without row exchanges (WR^+ condition). In this case it is possible to apply Theorem 3.1 of [1] (see (12)-(14)). In addition, if A is an ASSR matrix with $\varepsilon_1 = 1$, then by Theorem 1, we can assure that, for all $t \in \{1, \ldots, n\}$, $\varepsilon_1(A) = \varepsilon_1(\widetilde{A}^{(t)}) = 1$. Thus, we can consider Theorem 4.1 of [1] for to obtain the bound of $|E|$. If $\varepsilon_2(A) = -1$, we applied the result to to $A^{(1)} = P_n A$.

Acknowledgements

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

References

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

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

[3] P. Alonso, J.M. Peña and M.L. Serrano, *Almost strictly sign regular matrices and Neville elimination with two-determinant pivoting*, Preprint (2016) .

[4] T. Ando, *Total 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. **27** (1977) 485–490.

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

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

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

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

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

[11] R. Huang, *Componentwise backward error analysis of Neville elimination*, Linear Algebra Appl. **451** (2014) 33–48.

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

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

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

[15] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Real-Time Audio-to-Score Alignment using Multi-core Architectures

P. Alonso[1], P. Vera-Candeas[2], R. Cortina[3], F.J. Rodríguez-Serrano[2], M. Alonso-González[3] and J. Ranilla[3]

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

[2] *Telecommunication Engineering Department, Universidad de Jaén*

[3] *Depto. de Informática, Universidad de Oviedo*

emails: palonso@upv.es, pvera@ujaen.es, raquel@uniovi.es, fjrodrig@ujaen.es, monica300876@gmail.com, ranilla@uniovi.es

Abstract

The audio-to-score framework consists of two separate stages: preprocessing and alignment. In this work, we implement and evaluate an efficient real-time version of the alignment stage. Our program is based on a low complexity spectral decomposition for feature extraction and an online version of Dynamic Time Warping (DTW) for matching the score position with each input signal frame. The current version of the application has been tested on multi-core architectures including x86/x64 and ARM processor types. Experiments show that our framework allows to align in real-time scores of more than 30 minutes of duration on powerful systems (x86/x64 architecture), and scores about 15 minutes of duration on mobile devices (ARM architecture).

Key words: audio-to-score alignment, concurrent computing, real-time computing

1 Introduction

Interactive Music Systems encompass the use of real-time technology in music composition and live performance. Often this "interactivity" is associated with an electronic equipment driven by software that "listens" and "responds" to the performer's input in real-time. In the particular case of score-driven interaction, the core of this "intelligence" is the *score following*. In this work, we address the problem of audio-to-score alignment (or score matching),

which is the task of synchronizing an audio recording of a musical piece with the corresponding symbolic score [1]. There exist two approaches to this problem namely the so-called "offline" alignment and the called "online" alignment. Online alignment, also known as score following, processes the data as the signal is acquired. This tracking is very useful for applications such as automatic page turning [2] or automated computer accompaniment of a live soloist [3]. Other possible applications include synchronization of live sound processing algorithms for instrumental electroacoustic composition or the control of visual effects synchronized with the music (e.g. stage lights or opera supertitles).

Audio-to-score alignment is traditionally performed in two stages: *preprocessing* and *alignment*. First, the preprocessing stage is devoted to learn information from the score. Then, the alignment stage is generally subdivided into two steps: *feature extraction* and *matching*. The features extracted from the audio signal characterize some specific information about the musical content. Different representations of the audio frame have been used such as the output of a short-time Fourier transform (STFT) [4], auditory filter bank responses [5], or multi-pitch analysis information [6, 7]. Finally, the matching is performed by finding the best match between the feature sequence and the score. Two methods well known in speech recognition have been extensively used in the literature: statistical approaches (e.g. HMMs) [6, 8–10], and Dynamic Time Warping (DTW) [11–14].

In this work we present an efficient implementation of the alignment stage in real-time. Our implementation is mainly based on the approach proposed in [11] and uses the parallel online DTW solution presented in [15]. We have made an important effort in the implementation of the feature extraction stage so that the whole score following system can run efficiently.

2 Score following method

As we have previously noted, the proposed framework for real-time audio-to-score alignment has two parts: offline preprocessing stage and online alignment stage. The aim of preprocessing stage is to adequately organize the information given by the score to be used for alignment purposes. The output of the preprocessing stage consists of the spectral patterns $\mathbf{B}(f, k)$ for the K different units learned in advance using a MIDI synthesizer and kept fixed. The online alignment stage starts estimating the *gain* matrix $\mathbf{G}(k, t)$ and the *cost* matrix $\mathbf{D}(\tau, t)$, which measures the suitability of each unit to be active at each frame t (referenced to the signal input) by analyzing the likelihood between the spectral patterns $\mathbf{B}(f, k)$ and the input signal spectrogram. One component per unit is used so that, single non-zero restriction could be imposed to the gains allowing the use of the efficient signal decomposition method described in [16], among others. Further details can be obtained in [16] and [17].

Algorithm 1 summarizes the main steps to obtain the distortion matrix $\mathbf{\Phi}(k, t)$, where

T denotes the number of real signal frames. To obtain the optimum unit at each frame we use the Beta-divergence function as a cost function (see Eq. (1)), witch includes in its definition the most used costs functions in the state-of-art [16],

$$D_\beta(x|\hat{x}) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)\hat{x}^\beta - \beta x\hat{x}^{\beta-1}\right) & \beta \in (0,1)\cup(1,2] \\ x\log\frac{x}{\hat{x}} - x + \hat{x} & \beta = 1 \\ \frac{x}{\hat{x}} + \log\frac{x}{\hat{x}} - 1 & \beta = 0\ . \end{cases} \quad (1)$$

Algorithm 1 Distortion matrix computation method

```
1 Initialize B(f,k) with the values learned in preprocessing.
2 for t=1 to T do
3   for k=1 to K do
4     Compute the gains g_{k,t} using g_{k,t} = (sum_f x_t(f) b_k(f)^{(β-1)}) / (sum_f b_k(f)^β)
5     Compute the distortion values: Φ(k,t) = D_β(x_t(f) | g_{k,t} b_k(f))
6   end for
7 end for
```

Once the feature extraction step is completed, that is, the distortion matrix has been obtained, the cost matrix $\mathbf{D}(\tau, t)$ must be computed in order to perform the alignment, matching the score position with each input signal frame. Note that, in matrix $\mathbf{D}(\tau, t)$, τ represents the synthetic signal frames or score positions. To do this, we use the online version of Dynamic Time Warping (DTW) algorithm proposed in [11], with a fixed latency of just one frame, that was implemented in parallel in [15]. In order to obtain this low latency no backtracking is allowed, that is, the decision is made directly from the information at each frame t. As explained in [17], the simplest online approach is obtained by matching the performance position at frame t with the score position τ associated to the minimum value of the accumulated cost at frame t.

3 Evaluation and experimental results

As we previously stated, the aim of this work is to develop a high performance online score matching software. The implemented software makes use of double precision real numbers (64 bits). All the equations are directly implemented as shown in this work without any variation to reduce the complexity.

We carried out experiments in order to know the limits of a real-time response by the implemented score matching algorithm. The complexity per frame of the algorithm depends on the length of the score, thereby, some synthetic audio files whose duration varies from 150 seconds to 1800 seconds were used as benchmark on two different multi-core architectures.

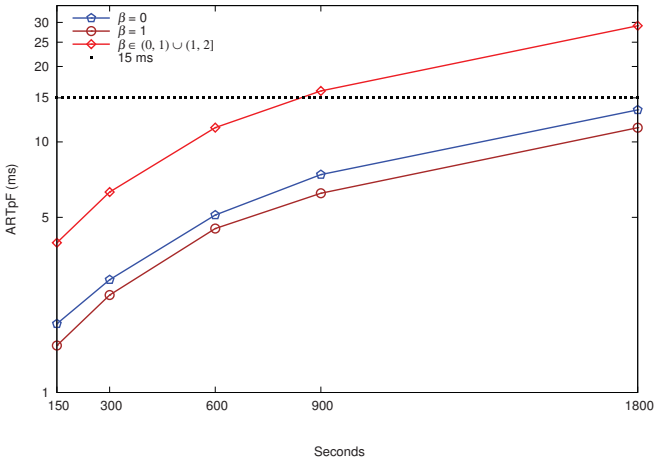


Figure 1: Evolution of the ARTpS using x86/x64 architecture

In conclusion, the experiments were designed to estimate the average run time per frame of the score matching algorithm as a function of the duration of the score. All the parameters of the score alignment algorithm were set to the same values given in [17].

As a representative of the x86/x64 architecture we used a server with one Intel® Xeon® E5-5620 @ $2.40GHz$ processor with 4 cores. Figure 1 shows the AveRage Time per Sequence (ARTpS) in this machine. The flat line that intersects the y-axis at $15ms$ corresponds to the maximum acceptable value for the ARTpS to be below the real-time threshold (with the hop size between frames used in [17]).

Notwithstanding the value of β does not have a strong impact on the alignment results (see [17]), different values of β may imply the use of the *log* function, only multiplications, or power functions (see Eq. 1) so, consequently, may highly impact on the ARTpS. We can see in Fig. 1 that the implemented score matching is able to be executed real-time on the x86/x64 architecture for those scores under 1800 seconds with β equal to 0 or 1.

Regarding the ARM® architecture, we used a system with a Cortex® A15 processor @ $2.32GHz$ with 4 cores. As in the x86/x64 case this device runs Linux OS. We show in Fig. 2 the ARM® architecture AveRage Time per Sequence (ARTpS). As it can be seen, the implemented score matching is able to execute in real-time only with small scores for all the admitted β values. When β is 0 or 1 the system responds in real-time for scores below ≈ 900 seconds.

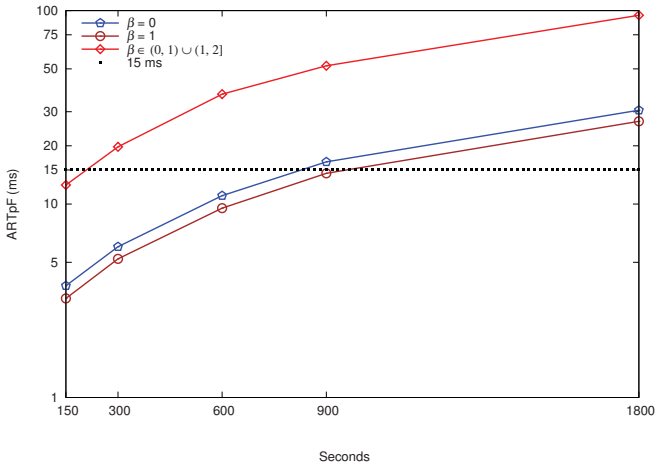


Figure 2: Evolution of the ARTpS using ARM® architecture

4 Conclusions and future work

In this work we have presented a useful tool for the Audio-to-Score alignment problem. Our proposal incorporates high performance computing techniques with the aim of performing the score following in real-time, which is a critical requirement of many applications. The current version has been implemented for multi-core architectures, encompassing x86/x64 processors and ARM® processors. To demonstrate the real-time execution of our implementation, we have performed experiments that analyze the limits of a real-time response in terms of score duration. This framework achieves good score alignments within the real-time constraint of $15ms$.

As a future work, we plan to do some algebraic transformations in the equations in order to reduce the total complexity. Also, we plan to expand our proposal to include coprocessors such as GPUs. As a result, we hope we can do real-time score following of longer audio files, and even reduce the real-time threshold below $15ms$ to tackle a larger set of applications.

Acknowledgments

This work was supported by the Ministry of Economy and Competitiveness from Spain (FEDER) under projects TEC2015-67387-C4-1-R, TEC2015-67387-C4-2-R and TEC2015-67387-C4-3-R, the Andalusian Business, Science and Innovation Council under project P2010-TIC-6762 (FEDER), and the Generalitat Valenciana PROMETEOII/2014/003.

References

[1] A. Cont, D. Schwarz, N. Schnell, and C. Raphael. 2007. "Evaluation of real- time audio-to-score alignment" *in Proc. of the International Conference on Music Information Retrieval (ISMIR) 2007*, Vienna, Austria.

[2] A. Arzt. 2008. "Score Following with Dynamic Time Warping. An Automatic Page-Turner" *Master's Thesis, Vienna University of Technology*, Vienna, Austria.

[3] C. Raphael. 2010. "Music Plus One and Machine Learning" *in Proc. of the 27 th International Conference on Machine Learning*, Haifa, Israel, 21-28.

[4] A. Cont. 2010. "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 974-987.

[5] N. Montecchio, and N. Orio. 2009. "A discrete filterbank approach to audio to score matching for score following," *in Proc. of the International Conference on Music Information Retrieval (ISMIR)*, pp. 495-500.

[6] Z. Duan, and B. Pardo. 2011. "Soundprism: An Online System for Score-informed Source Separation of Music Audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205-1215.

[7] M. Puckette. 1995. "Score following using the sung voice," *in Proc. of the International Computer Music Conference (ICMC)*, pp. 175-178.

[8] A. Cont. 2006. "Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical hmms," *In Proc. of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Toulouse. France.

[9] P. Cuvillier, and A. Cont. 2014."Coherent time modeling of Semi-Markov models with application to realtime audio-to-score alignment". *In Proc. of the 2014 IEEE International Workshop on Machine Learning for Signal Processing*, 16.

[10] C. Joder, S. Essid, and G. Richard. 2013."Learning optimal features for polyphonic audio-to-score alignment." *IEEE Transactions on Audio, Speech, and Language Processing*, 21, 10, 2118-2128.

[11] J.J. Carabias-Ortí, F.J. Rodríguez-Serrano, P. Vera-Candeas, N.Ruiz-Reyes, F.J. Cañadas-Quesada. 2015. "An Audio To Score Alignment Framework Using Spectral Factorization And Dynamic Time Warping," *in Proc. of the International Conference on Music Information Retrieval (ISMIR)*, pp. 742-748, Málaga, Spain.

[12] S. Dixon. 2005. "Live tracking of musical performances using on-line time warping," *in Proc. International Conference on Digital Audio Effects (DAFx)*, Madrid, Spain, pp. 92-97.

[13] N. Hu, R. B. Dannenberg, and G. Tzanetakis. 2009 "Polyphonic audio matching and alignment for music retrieval," *in Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 185-188.

[14] N. Orio, and D. Schwarz. 2001. "Alignment of monophonic and polyphonic music to a score," *in Proc. International Computer Music Conference (ICMC)*.

[15] P. Alonso, R. Cortina, F. J. Rodríguez-Serrano, P. Vera-Candeas, M. Alonso-Gonzalez, and J. Ranilla. 2016. "Parallel online time warping for real-time audio-to-score alignment in multi-core systems," *The Journal of Supercomputing*, published online DOI 10.1007/s11227-016-1647-5.

[16] J.J. Carabias-Ortí, F.J. Rodríguez-Serrano, P. Vera-Candeas, F.J. Cañadas-Quesada, and N. Ruiz-Reyes. 2013 "Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription," *Engineering Applications of Artificial Intelligence*, Volume 26, Issue 7, August 2013, pp. 1671-1680.

[17] J.J. Carabias-Ortí, F.J. Rodríguez-Serrano, P. Vera-Candeas, and D. Martínez-Muñoz. 2016. "Tempo Driven Audio-to-Score Alignment using Spectral Decomposition and Online Dynamic Time Warping," *ACM Trans. Embedd. Comput. Syst.*, accepted.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A technique to avoid order reduction in the integration of linear initial boundary value problems with Lie-Trotter method

Isaas Alonso-Mallo[1], Begoña Cano[1] and Nuria Reguera[2]

[1] *IMUVA, Departamento de Matemática Aplicada, Universidad de Valladolid, Spain*

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

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

Abstract

In this work, we show how to avoid order reduction when integrating linear initial boundary value problems with a classical exponential splitting method: Lie-Trotter method. Numerical experiments corroborate that order reduction is completely avoided when using the technique proposed in this work.

Key words: Exponential splitting methods, linear initial boundary value problems, time order reduction
MSC 2000: 65M12 65M20

1 Introduction

Splitting methods are frequently used in the literature for integrating differential equations. These methods are interesting when the numerical integration of separated parts of the equation is cheaper or easier than the numerical integration of the whole equation. In this work we focus on an exponential splitting method of first order: Lie-Trotter method. In [5], this method is analyzed when integrating linear initial boundary value parabolic problems with homogeneous Dirichlet boundary conditions. For Lie-Trotter, it is shown order reduction to one for the local error although no order reduction appears for the global error. With non-homogeneous boundary conditions, the order reduction is even bigger.

In this work we propose a generalization of Lie-Trotter method which does not show order reduction even for problems with non-homogeneous and time-dependent boundary

 ISBN: 978-84-608-6082-2

conditions. Moreover, the computational cost of this new technique is negligible compared with the whole cost of the method. Theoretical results and an analogous study for Strang method can be found in [4]. On the other hand, similar techniques have been proposed for other type of methods in [1, 2, 3].

2 Exponential Lie-Trotter splitting

Let X and Y be Banach spaces and let $L : D(L) \to X$ and $\partial : D(L) \to Y$ be linear operators. We are going to consider full discretizations of the linear abstract non homogeneous initial boundary value problem

$$\begin{array}{rcl} u'(t) & = & Lu(t) + f(t), \quad 0 \leq t \leq T, \\ u(0) & = & u_0 \in X, \\ \partial u(t) & = & g(t) \in Y, \quad 0 \leq t \leq T. \end{array} \tag{1}$$

Moreover, we assume that

$$L = A + B, \tag{2}$$

where $A : D(A) \to X$ and $B : D(B) \to X$ are linear operators which are supposed to be simpler than L in some sense, and $D(L) \subseteq D(A) \cap D(B)$.

In order to generalize Lie-Trotter exponential method to the case of non-vanishing boundary conditions, we suggest to advance a stepsize from u_n to u_{n+1} considering first the solution of

$$\begin{array}{rcl} v_n'(s) & = & Bv_n(s), \\ v_n(0) & = & u_n + kf(t_n), \\ \partial_B v_n(s) & = & \partial_B \left[u(t_n) + kf(t_n) + sBu(t_n) \right]. \end{array} \tag{3}$$

Secondly, we take the solution of

$$\begin{array}{rcl} w_n'(s) & = & Aw_n(s), \\ w_n(0) & = & v_n(k), \\ \partial_A w_n(s) & = & \partial_A \left[u(t_n) + kf(t_n) + kBu(t_n) + sAu(t_n) \right]. \end{array} \tag{4}$$

Then, we define u_{n+1} as

$$u_{n+1} = w_n(k). \tag{5}$$

3 Examples and numerical experiments

Now, we are going to integrate a parabolic problem with a dimension splitting and we are going to observe that, with the technique that we propose, order reduction is avoided.

Let us assume a and b are sufficiently smooth positive coefficients that are bounded away from zero, and let us consider the following parabolic problem, defined on $0 \leq x, y \leq 1$, $0 \leq t \leq T$,

$$\begin{aligned} u_t(t,x,y) &= (a(x,y)u_x(t,x,y))_x + (b(x,y)u_y(t,x,y))_y + f(t,x,y), \\ u(0,x,y) &= u_0(x,y) \\ u(t,0,y) &= g_{1,0}(t,y), \\ u(t,1,y) &= g_{1,1}(t,y), \\ u(t,x,0) &= g_{2,0}(t,x), \\ u(t,x,1) &= g_{2,1}(t,x). \end{aligned} \tag{6}$$

For this problem we use the splitting

$$A = D_x(a(x,y)D_x), \quad B = D_y(b(x,y)D_y).$$

We first consider the spatial discretization of problem (3), obtaining

$$\begin{aligned} \dot{v}_h(s) &= B_{h,0}v_h + C_h(\partial_B u(t_n) + s\partial_B Bu(t_n) + k\partial_B f(t_n)) \\ v_h(0) &= P_h(u_h + kf(t_n)) \end{aligned} \tag{7}$$

where $P_h(u_n + kf(t_n))$ represents the interior nodal values and $B_{h,0}$ is a block-diagonal matrix whose base matrix for each $m \in \{1, \ldots, N-1\}$ is given by

$$\frac{1}{h^2}\text{tridiag}(b(x_m, y_{l-\frac{1}{2}}), -(b(x_m, y_{l-\frac{1}{2}}) + b(x_m, y_{l+\frac{1}{2}})), b(x_m, y_{l+\frac{1}{2}})),$$

with $y_{l-\frac{1}{2}} = (y_{l-1} + y_l)/2$. Moreover, $C_h\partial_B u$ is a block-vector where each m-block has the form

$$\frac{1}{h^2}\begin{bmatrix} b(x_m, y_{\frac{1}{2}})u(x_m, y_0) \\ 0 \\ \vdots \\ 0 \\ b(x_m, y_{N-\frac{1}{2}})u(x_m, y_N) \end{bmatrix}.$$

Solving now problem (7), we obtain

$$\begin{aligned} \hat{v}_h &= e^{kB_{h,0}}P_h(u_n + kf(t_n)) + \int_0^k e^{(k-s)B_{h,0}}C_h\left[\partial_B u(t_n) + s\partial_B Bu(t_n) + k\partial_B f(t_n)\right] ds \\ &= e^{kB_{h,0}}P_h(u_n + kf(t_n)) + k\varphi_1(kB_{h,0})C_h\left[\partial_B u(t_n) + k\partial_B f(t_n)\right] \\ &\quad + k^2\varphi_2(kB_{h,0})C_h\partial_B Bu(t_n) \end{aligned}$$

 ISBN: 978-84-608-6082-2

where φ_1, φ_2 are the standard functions which are used in exponential methods.

Now, we consider the spatial discretization of problem (4),

$$\begin{aligned} \dot{w}_h(s) &= A_{h,0}w_h + C_h(\partial_A u(t_n) + k\partial_A Bu(t_n) + k\partial_A f(t_n) + s\partial_A Au(t_n)) \\ w_h(0) &= \hat{v}_h, \end{aligned} \tag{8}$$

ans we obtain

$$\begin{aligned} \hat{w}_h &= e^{kA_{h,0}}\hat{v}_h + \int_0^k e^{(k-s)A_{h,0}} C_h \left[\partial_A u(t_n) + k\partial_A Bu(t_n) + k\partial_A f(t_n) + s\partial_A Au(t_n)\right] ds \\ &= e^{kA_{h,0}}\hat{v}_h + k\varphi_1(kA_{h,0})C_h \left[\partial_A u(t_n) + k\partial_A Bu(t_n) + k\partial_A f(t_n)\right] \\ &\quad + k^2\varphi_2(kA_{h,0})C_h\partial_A Au(t_n). \end{aligned}$$

Then, we define $u_{n+1,h}$ as

$$u_{n+1,h} = \hat{w}_h.$$

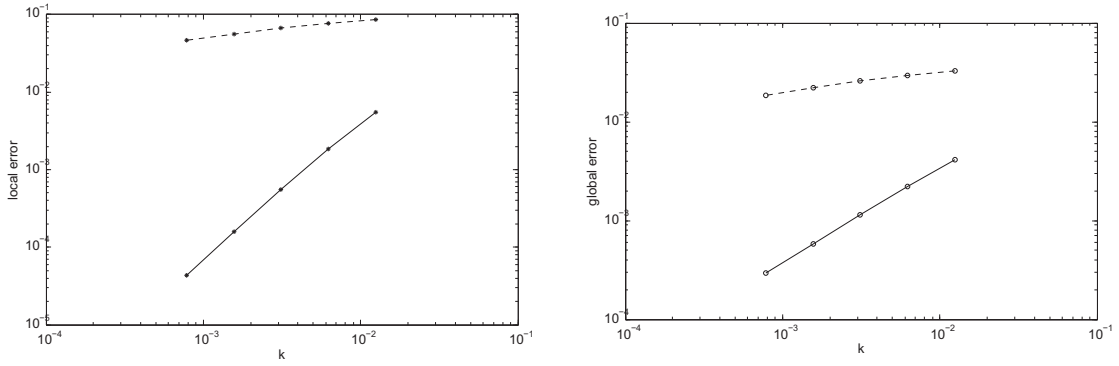


Figure 1: Errors without avoiding (discont.) and avoiding (cont.) order reduction when integrating problem (6) with Lie-Trotter method.

Let us show, with a numerical experiment, that with the implementation we propose for Lie-Trotter method, we avoid order reduction. More precisely, we are going to integrate problem (6) with

$$\begin{aligned} a(x,y) &= 1 + x + y, \\ b(x,y) &= 1 + 2x + 3y, \\ u_0(x,y) &= (x^2 - 1/4)(y^2 - 1/4), \\ f(x,y,t) &= \frac{e^{-t}}{16}(15 + 56y - 28y^2 - 32y^3 + x(32 - 64y^2) - 4x^2(7 + 48y + 4y^2) - 64x^3). \end{aligned}$$

The exact solution of this problem is $u(x,y,t) = e^{-t}(x^2-1/4)(y^2-1/4)$. In Figure 1 we show the local and global errors when we integrate the problem without avoiding order reduction and with the technique proposed in this work with $h = 10^{-2}$. We observe that both for the local and global errors, the results without avoiding order reduction are very poor in terms of the order and of the size of the errors. Nevertheless, if we apply the technique proposed in this work, we avoid order reduction, observing order 2 and 1 for the local and global errors respectively and the errors are much smaller. This numerical experiment corroborates the theoretical results which are proved in [4].

Acknowledgements

This work has been partially supported by MTM2015-63837-P.

References

[1] I. Alonso–Mallo, *Runge-Kutta methods without order reduction for linear initial boundary value problems*, Numerische Mathematik **91** (2002), pp. 577–603.

[2] I. Alonso–Mallo, B. Cano and J. C. Jorge, *Spectral-Fractional Step Runge-Kutta Discretizations for Initial Boundary Value Problems with Time-Dependent Boundary Conditions*, Math. Comput. **73** (2004), pp. 1801–1825.

[3] I. Alonso–Mallo, B. Cano and N. Reguera, *Avoiding order reduction when integrating linear initial boundary value problems with Lawson methods*, submitted for publication in 01/2015, available at http://hermite.mac.cie.uva.es/bego/RAhd.html

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

[5] E. Faou, A. Ostermann and K. Schratz, *Analysis of exponential splitting methods for inhomogeneous parabolic equations*, IMA J. Numer. Anal. **35** (1) (2015), 161–178.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A secure and efficient ecc based method to avoid impersonation for the SIP protocol.

J.A. Alvarez-Bermejo[1] and J.A. Lopez-Ramos[2]

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

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

emails: jaberme@ual.es, jlopez@ual.es

Abstract

The session initiation protocol (SIP) has been selected as the official end-to-end signallig protocol for establishing multimedia sessions in the IP-based Universal Mobile Telecommunication Systems (UMTS) network. Supporting session establishment even in S-UMTS (satellite linked networks) is a requirement. Providing secure and fast methods is a mandatory task. SIP is a powerful signalling protocol for trasnmitting media over internet protocol. Authentication is an important security requirement for SIP. Hitherto, many authentication schemes have been proposed to enhance the security of SIP. Being the impersonation problem one of the hot topics. In this paper, a novel authentication and key agreement scheme is proposed for SIP using elliptic curve cryptosystem (ECC). Security and performance analyses demonstrate that the proposed scheme is secure against security attacks of various types and has low computation cost and a low energy profile.

Key words: Elliptic curve cryptosystem, Impersonation attack protection, Session initiation protocol

1 Introduction

In the last decade, the growth of the Internet has escaled beyond anyone's reasonable imagination into a universal communication platform. Running almost concurrently with this Internet proliferation has been the equally extraordinary escalation in the number of mobile wireless networks and subscribers. The emergence of IP-based networks for mobile communications will finally enable of, and the access to voice, video, messaging, data and

web-based technologies for the wireless user in a seamless manner, in much the same way that fixed users are currently experiencing over the Internet. The provision of IP-based multimedia services in UMTS is made possible through the introduction of the IP Multimedia Subsystem (IMS) as part of the Third Generation Partnership Project (3GPP) release 5 set of standards. Centred around the provision of IP multimedia services is the end-to-end signalling based on the Session Initiation Protocol [1], a protocol developed within the Internet Engineering Task Force (IETF). SIP has been selected by the 3GPP as the official end-to-end signalling protocol for establishing multimedia sessions in UMTS [2] so as to be compliant with the rest of Internet to ease interoperability, as well as due to the many positive attributes SIP posseses, such as its simplicity, extensibility, flexibility, and scalability.

The session initiation protocol (SIP) is an application layer protocol which is used as a signaling one for establishing, altering, and terminating multimedia sessions between different users in IP based telephony systems. One of the most important aspects with SIP is authentication of the requester of the service, clients need to know the identity of a person that is peering the multimedia communication, also service operators need to identify the requester in order to charge, and provide preferential services or differential treatments. So the person who claims access to the service must be unambiguosly identified. And in addition to this, the system must be able to detect the missuse of credentials due or a double spend.

An efficient and secure authentication and key agreement scheme is able to provide various aspects of security for SIP communications. With an authentication and key agreement scheme, the client and the server can mutually authenticate each other, they can negotiate a session key for secure communication.

This negotiated session key can be used to give access to a decyphered communication, for example. User and server, both can get to use a symmetric encryption algorithm to get confidentiality. Services and solutions to provide integrity and detect unauthorized modifications are provided by a number of algorithms. But, the computational overhead of such services may harm transmission delays, which is not acceptable for the real-time communications [3], the scenario gets even worst when the S-UMTS is involved.

With the increasing use of applications that create softphones on laptops and smartphones, security concerns have also increased as the mobile devices are more likely to be stolen or lost. Given these devices, the security to access the service must be rethinked. There are attacks such as the cold boot attack (see [4]) that permits getting the keys from RAM using elementos to freeze integrated circuits to retain data in memory mean while the data is dumped into a secondary device. Therefore time stamp marks, and crypto counters are needed in order to guarantee that the user re-authenticates so often. There are approaches, in this sense, like the one described in [5] that showed that continuously recorded touch data from a touchscreen and use it as a behavioural biometric pattern able

to authenticate the user. This process is interesting but is not really practical as the system is conducted to take additional operations that are not power aware.

Authentication is usually made through a PKI infrastructure using classical certificates. However identity is revealed through the authentication process and for some applications anonymity would be desirable. And this can be done by means of a *licence* as we propose in this paper. Group authentication based in for many-to-many authentication is good for audio and video multiconference in VoIP. Nevertheless, communication schemes in SIP are a one-to-one. In this scenario, the HTTP Digest authentication method is what is considered as normal for SIP. But it does not give mutual authentication and is, also, insecure against offline password guessing, etc. SIP, therefore, is a protocolo where researchers are investing their efforts to provide authentication schemes, some of the are [6], [7], [8].

Although there are plenty of proposals, many of them fails in keeping good security levels. For example, [6] proposal is built upon the difficulty of the Discrete Logarithm Problem (DLP); therefore, considered as secure but attacks like the one shown in [7] shows that the protocol fails. In fact, this paper ([6]) as it is built on DLP has been target for many other attacks. The protocol we propose, resist all these attacks and even offer a mean to guarantee the identity of the user when accessing a service.

Taking DLP as a feasible path to create almost secure schemes, in [9], authors did a authentication scheme based in ECDLP for SIP. Being its main advantage the fact that the elliptic curve cryptosystems (ECC) needs a smaller key size compared to the other cryptosystems in order to provide the same security level. This was considered as a notable improvement on [6] approach. But also was attacked showing its weakness. Many other schemes proposed, all, have been attacked and shown to be vulnerable to offline password guessing, Denning-Sacco attacks, impersonation attacks. Many other protocols are all vulnerable to privileged insider attacks. Because SIP accepts plain text, in the registration stage of these protocols the user send the selected password data in plain to the server. So the privileged insider can impersonate the user in many services.

A feasible solution to these is the one proposed by [8] for securing SIP based in the construction of a smart-card-based authentication and key agreement scheme. Here the smartcard is used as a second authentication factor. The smartcard is required to further proceed during the access to services requested by the user. This protocol is vulnerable to impersonation attacks. In this paper, we propose the usage of a licence which is built containing information from a ticket T built collaborativelly between user and server. Solving the impersonation attack problem. The proposed scheme not only can improve the security, but also is to be more efficient than the previous authentication schemes proposed for SIP.

The rest of this paper is organized as follows. Section 2, provides a description of the proposed protocol. The security proofs are provided in Section 3. Section 4, presents tests where the efficiency of the protocol is shown. Finally, Section 5 concludes the paper.

 ISBN: 978-84-608-6082-2

2 Protocol

We introduce a protocol based on blind signatures using elliptic curves that allows to trace those users of any service,traitors, that share their legitimate license (or that it has been stolen) with other people, avoiding the abuse or the unauthorized use of legal licenses. We show that our protocol is incapable of implicating innocent users in the illegal distribution of licenses and that tracing produces indisputable proof for the implication of the traitors. We also show that our protocol provides anonymity to users unless they share their private information. We also provide an implementation showing its practical application.

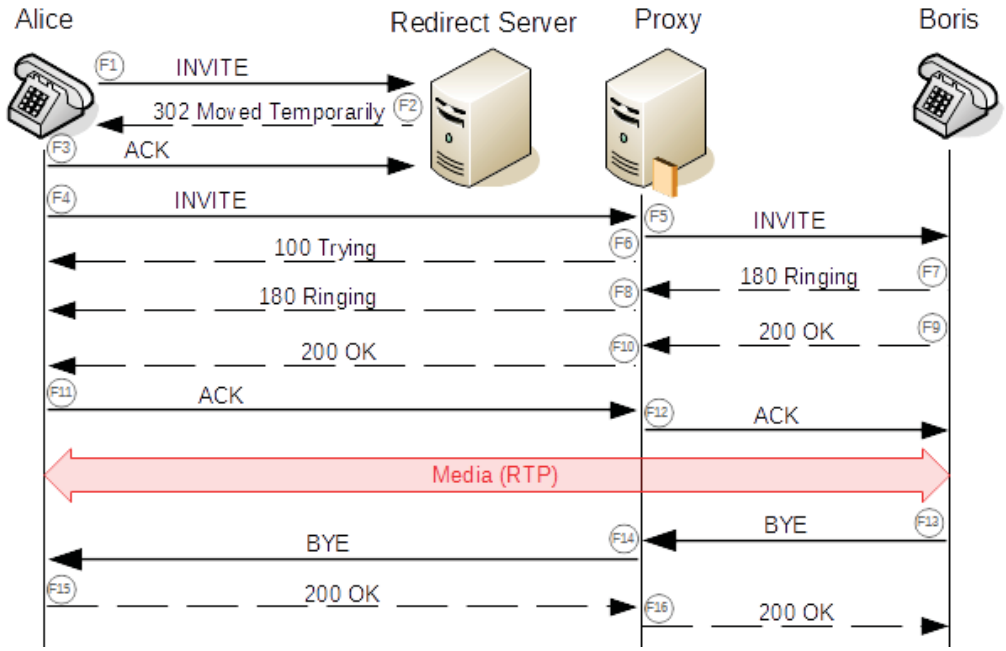


Figure 1: SIP call message flow exchange

The protocol we are introducing in this section deals with an on-line protected service provided by a content server that implements the SIP protocol, therefore extending it, that is susceptible of illegal use by users or outsiders. The service will be accessible by means of a session key, also distributed by the content server or by a keys server. This session key will be broadcasted in an encrypted way and users will recover it by using a private key, namely a ticket, that is previously distributed individually at the moment of the service subscription. The ticket will be accompanied by a unique license, that will be required every time a user demands access to the service. A third part that will be called the agent will check the validity of the license and in that case this will send it to the server or a

second agent that may detect its possible illegal use.

SIP is described in RFC 3261, controlled and ratified by the Internet Engineering Task Force (IETF), and describes a typical SIP call between the talkative Bob and Alice, containing a message sequence (see figure 1).

The protocol is composed of the following phases:

- Setup: during this stage, the SIP server choses an elliptic curve and a group of points from that curve. Creating internal and private keys. Using these private keys to expose public values built on these keys. Apart from this, the call or session initiation needs to achieve several distinct functions, and therefore the server must create the necessary resources, including:
  - Register a device for each user
  - Establish a data path between the two users
  - Negotiate a data format to use for voice data
  - Transmit /receive voice data (decrypted if the license is valid)
  - Terminate the call
- License generation

  Let us suppose that some user is to require services from the SIP server. During its first connection, the user must get registered in the server. In this point the user runs code to extract unique information from the hardware plaftorm, and verify that this number is valid. And keep it as part of the *identity* that must send to the server. The user gets the IMEI code for its hardware, that makes the device unique in the network. Each 3GPP mobile phone should have a unique identifier, the International Mobile Equipment Identity (IMEI) or Formula (IMEISV). The IMEI (15 decimal digits) or IMEISV (16 decimal digits). Currently, in wireless networks there is no functionality that prevents the use of counterfeit phones. Some countries are using the black list (EIR) to block mobile phones without a valid IMEI, but this method has proven to be ineffective for counterfeit and substandard phones, since an IMEI can be illegally changed (IMEI cloned). The first 8 of the 15 digits of an IMEI correspond to the type allocation code (TAC), which represents the phone model. The last digits correspond to the serial number and check digit (1 digit). In [10] a new method was developed in order to identify and block any type of fraudulent device, independent of the 3GPP technology. The novelty of the invention is cross-checking device capabilities with a database obtained from all vendors; it is simple, but strong and effective. The cross-checking procedure could occur in a mobile originating call (MOC), mobile termi-nating call (MTC), location update (LU), SMS, IMSI Attach, GPRS Attach, PDP context activation, RA update, service request, Short Network

Management Protocol (SNMP) queries, the operational system fingerprint (collected in the data link with core PS elements like SGSN or SGW), or any other procedure that could indicate a device's characteristics or capability. It needs to be configured on the network side since the IMEI, IMSI (user identity), and user equipment (UE) capabilities are collected and stored in the core network (CN), see figure 2.

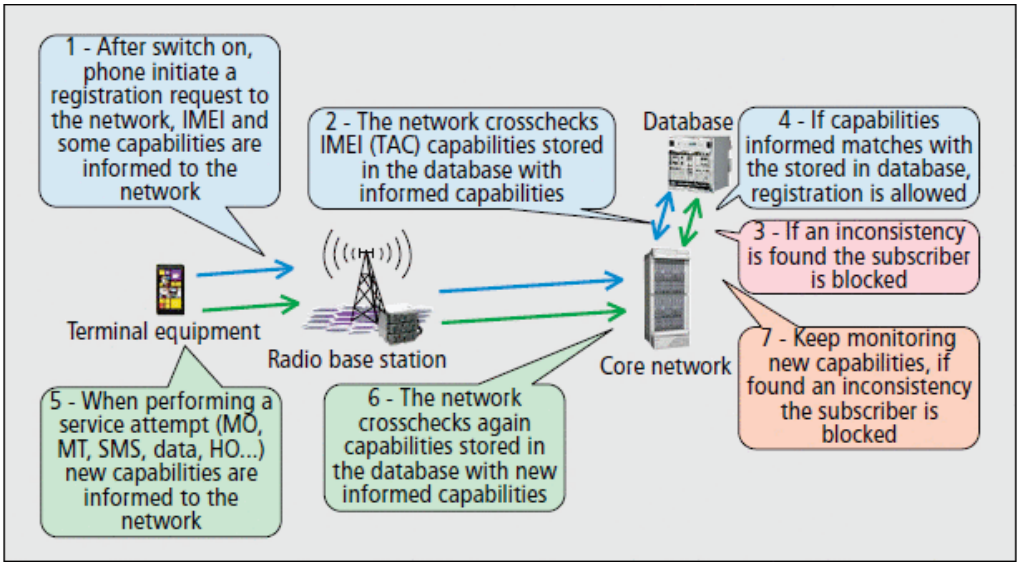


Figure 2: Ensure IMEI is unique

After the check is done, the user is assigned a valid, and unique IMEI code. That is dependent on the hardware that the smartphone has. As a second factor to create a number that relates user and device, a second procedure is included in the protocol. A MD5 fingerprint of the firmware loaded in ram is obtained. This number gets the reflect of the current hardware configuration as the firmware sees it. Any changes made to it would imply to renew the identity of the user. In other words, let think in this firmware as the BIOS of the system. When a BIOS boot it creates routines that permits the system boot (and the routines are hardware dependent). The number obtained by operating IMEI and BIOS MD5 with an XOR is set as the secret number u. The user selects a pasword (that identifies her in the device), this password is hashed as $\mathrm{H}(u||password)$. With it, the user creates a public identity $I = H(u||password)G_1$ that is submitted to the server. User also sends a message containing a tuple: username and password (or a 13 digit number).

The value computed by the user and sent to the server is used by the SIP extended protocol to create a Ticket that identifies the user (see figure 3).

 ISBN: 978-84-608-6082-2

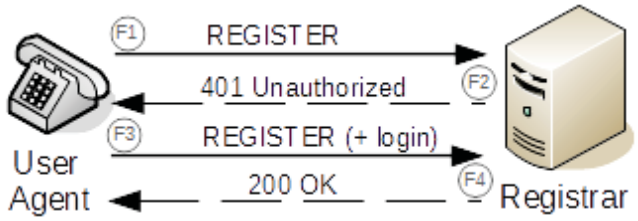


Figure 3: Extended registration message to add the ticket

- Requiring access to a service

  Alice decides to call Boris, so she enters Boris's address and initiates a call, causing Alice's UA to send a SIP INVITE message to its selected proxy. The challenge for the proxy is to obtain the IP address of Boris so that voice data can be routed between them. This is called the SIP discovery process. The discovery process was not extended.

  Once the INVITE message arrives at the correct proxy, it must be mapped to a specific IP address for Bori's SIP phone. The registration stage done by Boris was extended (as Alice's) to include specific information of the user so impersonation can be avoided. This is accomplished by looking up the URI in the location register. It will then return the current IP address registered during the *extended registration stage* by Boris. Now the INVITE message can finally be sent to the end device.

  Accepting the call is another challenge as it must be checked that both ends use the same data format, this is done by means of SDP (Session Description Protocol). Provided Boris' phone is the one that it claims to be, and that the data is compatible with one of these formats, Boris is notified that Alice is trying to call him. Once he answers the phone and accepts the call, his IP Address is sent back to Alice in a response message, along with an SDP payload containing the selected codec configuration, which then follows the same route back to Alice as the invite took on the way to Bob. She will then send an ACK message to signal the setup is complete.

  Now the access to the service is granted and the license can be used to determine when the access is granted or not.

3 Performance analysis

In this section, the performance of the proposed scheme is analyzed. First, we will analyze the impact of the proposed extension to the SIP protocol in terms of power consumption. To this we have emulated the smartphone architecture to conduct an energy footprint of

the device when registering and requesting a service. Creating a detailed power model is a quite complex task. Gem5 [13] is an architecture emulator that allows to create an ARM core (as the one that a conventional Android based smartphone has). And McPat [14] is an integrated power, area, and timing modeling framework that supports comprehensive design space exploration for multicore and manycore processor configurations ranging from 90 nm to 22 nm and beyond. The code run in Gem5 creates XML files that become the input for the McPAT emulator. This way we run our tests using the unextended SIP and the extended SIP versions. Figure 4 executes within the red box. If an average value is considered for the ease of use, we can state that the protocol is run and that the average consumption is around 0.4 mamperes. To test the protocol, the extended version was run several times. The figure 5 shows that the average consumption is also around 0.4mA being this a first proof that the extension implemented in SIP is almost negligible.

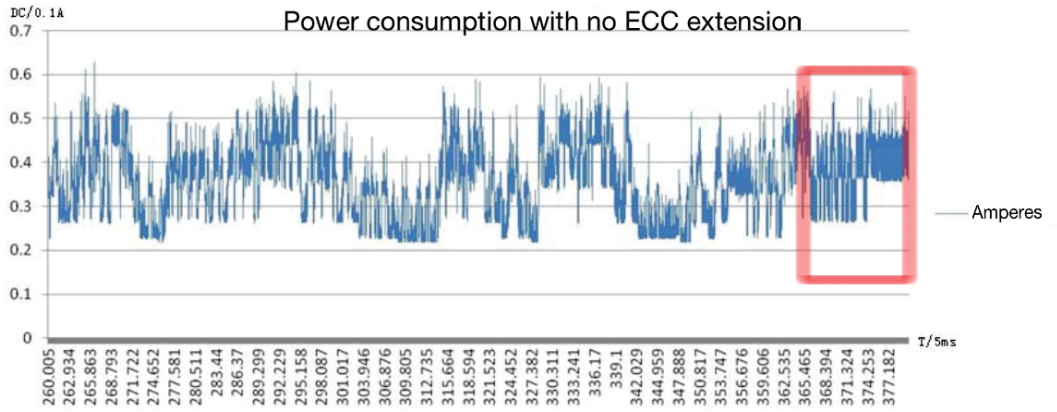


Figure 4: HTM Digest

3.1 Runtime

A mean time of 12,75864 milliseconds is needed in order to, first, register and later get granted access to any service to which we are connecting using the license. The conventional protocol with no extension to check identity and to avoid piracy runs in about 2 to 5 ms. Our protocol needs an average of 10 milliseconds to let the user in with additional security. A future line of research is opened trying to reduce this breach and make it smaller.

4 Conclusions

SIP is a protocol that uses a very simple architecture of layers to manage the initiation, and termination of sessions during which media is transmitted. The payments (micropayments)

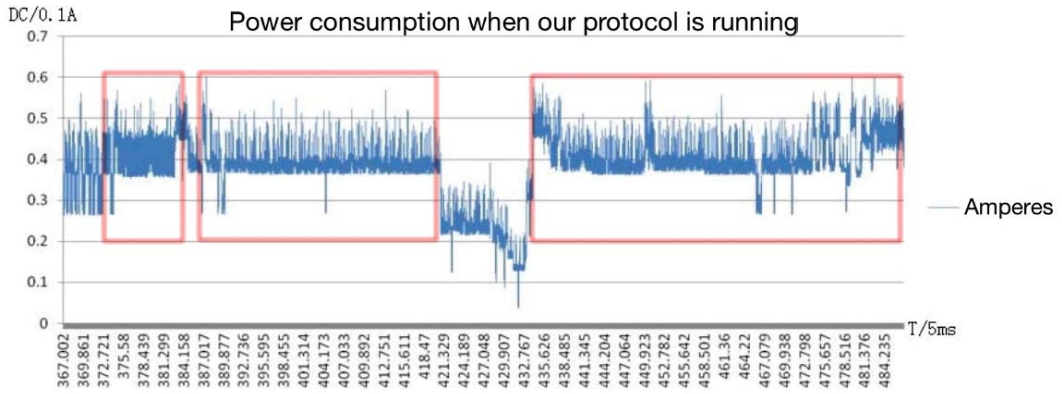


Figure 5: Licence control protocol

using such system is a topic in today's online services. Extending the protocol to create a secure instance of it, is a challenge as there are many attacks to observe. We have presented a method based in digital cash, extended with a licensing method to identify the user. This method has proved to be efficient and robust when facing the most common attacks, including the impersonation attack.

Acknowledgements

References

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, *SIP: Session Initiation Protocol*, IETF RFC 2002 **3261**.

[2] 3GPP TS 23.228 V5.70, *Architecture principles for release 2000*, Release 2000 **July 2000**.

[3] L. Zhou, H-C, Chao, AV. Vasilakos, *Joint forensics-scheduling strategy for delay-sensitive multimedia applications over heterogeneous networks.*, IEEE J Sel Areas Commun **29(7)** (2011) 1358–1367.

[4] B. Poettering, D. L. Sibborn, *Cold Boot Attacks in the Discrete Logarithm Setting.*, Topics in Cryptology — CT-RSA 2015: The Cryptographer's Track at the RSA Conference 2015, San Francisco, CA, USA, April 20-24, 2015. Proceedings **29(7)** (2015) 449–465.

[5] M. Frank, R. Biedert, E. Ma, I. Martinovic, D. Song , *Touchalytics: on the applicability of touchscreen input as a behavioral biometric for continuous authentication.*, IEEE Trans Inf Forensic Secur **8(1)** (2013) 136–148.

[6] C.C. Yang, R.C. Wang, W.T. Liu, *Secure authentication scheme for session initiation protocol*, Comput Secur **24** (2005) 381–386

[7] E.J. Yoon, K.Y. Yoo, C. Kim, Y. Hong, M. Jo, H. Chen, *A secure and efficient SIP authentication scheme for converged VoIP networks*, Comput Commun **33(14)** (2010) :1674–1681

[8] L. Zhang, S. Tang, Z. Cai, *Efficient and flexible password authenticated key agreement for voice over internet protocol session initiation protocol using smart card.* , Int J Commun Syst. (2013)

[9] A. Durlanik, I. Sogukpinar , *SIP authentication scheme using ECDH.*, World Enformatika Soc Trans Eng Comput Technol **8** (2005) :350-353

[10] A. J. Figueiredo Loureiro, D. Gallegos and G. Caldwell , *Substandard cell phones: impact on network quality and a new method to identify an unlicensed IMEI in the network.*, IEEE Communications Magazine **52(3)** (2014) :90-96

[11] C. Guang-Huei, C. Wen-Tsuen , *Secure broadcasting using the secure lock*, IEEE Transactions on Software Engineering **15(8)** (1989) :929-934

[12] B. Liu, W. Zhang, T. Jiang, *A scalable key distribution scheme for conditional access system in digital pay-TV system*, IEEE Transactions on Consumer Electronics **50(2)** (2004) :632-637

[13] F. A. Endo, D. Couroussé and H. P. Charles, *Micro-architectural simulation of in-order and out-of-order ARM microprocessors with gem5*, 2014 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS XIV) (2014) :266-273.

[14] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen and N. P. Jouppi, *McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures*, 2009 I42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO) (2009) :469-480.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Safe Control of Luggage with Homomorphic Cryptography

Néstor Álvarez-Díaz[1] **and Pino Caballero-Gil**[1]

[1] *Department of the Authors, Department of Computer Engineering and Systems*
Universidad de La Laguna

emails: nalvared@ull.edu.es, pcaballe@ull.edu.es

Abstract

The aim of this proposal is to provide passengers with a complete system that allows reducing the waiting time when people try to check-in their luggage as well as with an innovative tool for the control and the management of suitcases. This goal tries to cover the need of optimizing passengers' operations inside airport terminals. Besides, the proposal allows increasing the security of air transport from the point of view of the control of the luggage, which currently is critical due to the existence of terrorist threats. For this purpose, Near Field Communication (NFC) is used and wireless communications have been protected with a scheme based on homomorphic cryptography, and in particular, in the Paillier cryptosystem.

Key words: Airports, Check-in online, Luggage Control, NCF tags, Security

1 Introduction

Air transport is nowadays in a period of fast and great development with a huge number of requests that will provoke an unsustainable situation for the airports around the world in a short time. Therefore, the necessity to take actions in order to manage the luggage and the passengers as well as the security related to this process is absolutely essential. Nowadays, several tools and purposes for the optimization of the operations that the passengers have to do inside the airport terminals are a common topic [1]. The most innovative ones include changes in the check-in system or in the selection of the seats. In some reports about the passengers' satisfaction, the aspects that produce experiences inside the airport with more or less satisfaction level let us know different points to take in consideration. In particular, a study [2] has revealed a great change of passengers' attitudes in relation to the

availability of new technology inside the airport. Thus, luggage management (34%), real-time luggage control (40%) and auto-labelling of luggage (33%) are points that reflect a generalized dissatisfaction among passengers. The most common reason for these problems is the strengthening of the security policies in the air transport. These facilities are current objectives of terrorist attacks so security actions are necessary, and therefore they produce the long waiting times that the airport users suffer each day [3] [4].

In order to solve these problems, the main aim of this proposal is to provide a mechanism to let airport users realize common activities like luggage management, focusing on the increase of its control and providing the possibility of real-time luggage monitoring.

The general purpose of this work is to provide a system for the permanent luggage labelling through NFC tags in order to offer the possibility of identifying the owner as well as to realize an exhaustive tracking of each bag inside the airport, taking into account the safe requirements of this system and the guarantee of the authenticity and protection of data privacy and integrity. In particular, the proposal includes the use of the Paillier cryptosystem, which is a public-key algorithm with homomorphic properties. Besides, the current state of luggage could be known from everywhere thanks to the use of NFC tags. For the configuration and the reception of data only NFC readers are required. A current smartphones are beginning to include this technology so the passenger could manage the configuration data through a mobile phone application. Otherwise, when the users have not access to this technology in their smartphones, they will be offered the possibility to come to a check-in desk at the airport, which should be equipped with NFC technology. Thereby, similar to the current availability of online check-in, the passengers could check-in their luggage faster, in response to millions from passengers requests who every year suffer large waiting times for this operation.

This paper is structured as follows. Section 2 describes the general performance and the essentials aspects of the proposal. Also, in this section the NFC tags structure and the used fields are mentioned. The cryptographic scheme that provides reliability in order to control the luggage as well as protection of authenticity and confidentiality of data are described in Section 3. Possible vulnerabilities and attacks against to the system are analysed in Section 4 and also how they are solved. Finally, Section 5 closes the paper with some conclusions and open problems.

2 Overview of the Proposal

As aforementioned, the proposed system has two objectives that are tightly linked: the optimization of luggage check-in and the airport security in relation to the increase of luggage control [5]. For these purposes, a system to allow the identification of passengers and their luggage has been developed. The complete process is described as follows (see Fig. 1):

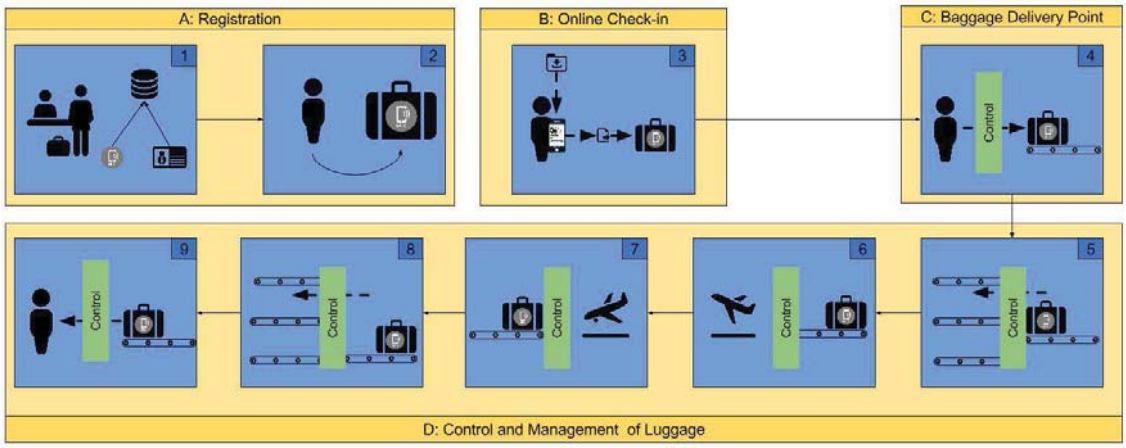


Figure 1: System Process

(A) Registration. The passenger goes to the airport and he or she asks to be the registered at the platform. Then, after checking his or her identity and storing in the system several basic data, the airport personnel gives him or her the required NFC tag. It should be noted that the same person can ask for more than one NFC tags as each tag identifies one bag and its relation with the owner, but the passenger could have several bags. After this first step, the system will always be able to know the identity of each person who is linked with each NFC tag. Then, the passenger can put permanently the NFC tag on a piece of his or her luggage and it is not necessary to ask for new tags in future trips with that luggage.

(B) Online Check-in. In the same web platforms or mobile phone applications that allow at passengers make their check-in online, they will be able to register their luggage. If a passenger does it, the system returns the encrypted and signed data of the trip so that this information can be written automatically in the NFC tag. Thus, the application notifies the user that he or she must put the device near the tag to make communication possible.

(C) Baggage Delivery Point. After configuring the NFC tag, the passenger can place the bags in the airport baggage delivery points, where the identification of the passenger as well as the authenticity and integrity of stored data in the tag are verified thanks to digital signatures.

(D) Control and Management of Luggage. Information about the process must be added at each Control point (C), from the delivery of the luggage in the delivery point to its

placement inside the plane and when the passenger arrives at the destination airport and he or she gets back the luggage.

In the present work, the NFC tag NTAG213 [6] has been chosen because it has 144 bytes for the user's data and it has a static serial Number (N). The space for these data is enough to store both the message and its signature, and the number N allows identifying unequivocally the owner of the tag and its relationship with the bag because this value is static. This aspect is interesting for the cryptographic system explained in Section 3. In relation with the registration, the N value allows linking in the server S the NFC tag with the user u and his or her bag j ($u = 1, 2, \ldots; j = 1, 2, ..., B_u$) where B_u is the passenger's total number of Bags whose owner is u.

The control points C_v ($v = 1, 2, \ldots, P$) where P is the total number of control points, are responsible for verifying that the information stored in the NFC tag has not been altered as well as for updating the current state of luggage in the NFC tag and server. A real-time tracking of luggage and the control of state of each bag is therefore be possible. For that, in each point C_v there is an NFC reader that allows reading and writing the NFC tag. Moreover, in the most critical points in relation with security, like delivery points, the presence of personnel who can verify the content of the luggage physically through X-rays for example, would be advisable. The system is composed of:

- **The server**, which allows managing the passengers and the luggage. This element acts as centre of all the system as it grants the check-in access for the passengers and their luggage, and the register of state changes for each bag and tracking issues. It should be noted that this server is endowed with the ability to encrypt and decrypt the NFC tags' content, being the unique entity with these permissions.
- **The smartphones** and airport check-in points, which provide the ability to write data on NFC tags. When the passengers check-in their bags, the NFC tags should be written in order to load the configuration data. These elements can only write on NFC tags and never possess the ability to read stored information.
- **The NFC tags**, which allow registering the set of states that the luggage has followed. The fact that the tag updating operation is available implies to take a more exhaustive control about the luggage. In the present systems and other proposals, the updating of information in the baggage tags is not possible due to the fact that nowadays used labels are printed bar codes and their substitution is impracticable.
- **The control points**, which are entities that verify the authenticity and integrity of the data written in the NFC tag and update them.
- **The verification authorities** located in some control points which have the function of deciding the normal cycle for each bag and of notification about an incorrect state if the situation requires it.

3 Cryptographic Security

Since the proposed system manages sensitive information, the privacy of data is essential. The tags with NFC technology can be read or written easily by a lot of free mobile applications. Therefore, the integrity, authenticity and confidentiality of stored data must be preserved. The set of operations executed during the process is shown in Figure 2 and they are described as follows:

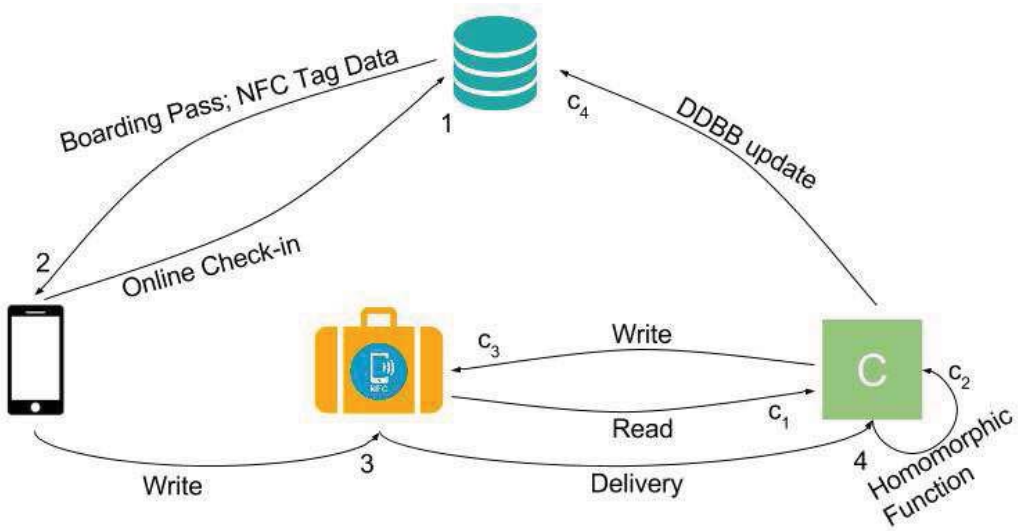


Figure 2: Data Transmission Scheme

- **The steps [1-4]** are the passengers' tasks that are executed after registration:

  1. *Online Check-in:* The passengers could obtain their boarding pass together with the data for writing in the NFC tag through the mobile phone application.
  2. *Boarding Pass and NFC Tag Data:* This step includes sending data from the server to the smartphone of the passenger. All data transmitted for being written into the NFC tag are sent encrypted and signed with the corresponding server keys. It should be noted that the server has two pairs of keys, one of them is for the encryption/decryption operation and the other one is for the digital signature.
  3. *Write Operation:* In this step the passenger sends the received information to the NFC tag through the mobile phone application.

4. *Delivery:* At this moment the passenger has written the information in the tag and goes to the delivery point to place his or her bag in it. The first control point C_1 is located at this platform and allows checking the identification of the integrity of written data and that the relation between the passenger and his or her luggage is correct.

- **The notation $\mathbf{c_k}; \mathbf{k} = \mathbf{1}, \mathbf{2}, ..., \mathbf{4}$** denotes four sets of operations that must be executed in each control point. These operations are executed after the luggage delivery and during all the process until the owner gets back their bags. Each set of operations is defined as follows:

  (c_1) *Reading NFC tag*: This operation is essential because it implies that the point C_v can read the tag information and obtain the serial number N and the encrypted and signed data. Then, the C_v verifies the signature with the pre-shared public key of the server S, which allows to check integrity and authenticity of the message. If the verification is not successful for that particular bag, it should be attended in a special way. On the contrary, the bag would follow the next steps.

  (c_2) *Homomorphic function*: A Partially Homomorphic Encryption (PHE) [7] is proposed in order to prevent a server overload with a huge demand of requests in each control point. The server S is the unique entity endowed with the ability to decrypt the information that is read in the previous step and that in each point C_v must be changed without an homomorphic function. For changing a minimum quantity of information, the point C_v would have to send to the server the encrypted data and the corresponding changes and then the server should reply with the new encrypted data and C_v write this information in the NFC tag. Thus, to avoid this situation, the use of the homomorphic cryptography has been chosen so it allows changing the content of a message without decrypting it previously. Therefore, without the necessity of interaction with the server S, each control point is able to change the content of the NFC tag allowing saving resources and network traffic.

  (c_3) *Writing NFC tag*: It should be noted that the message is composed by two parts: the own message and its signature. Firstly, the server signs the message with its keys. Secondly, in each control point the information changes and it is necessary that the point signs another the time new message with its keys. As shown in Figure 3, each control point should verify the signature of the previous point. Then the system can always check the integrity and the authenticity point by point. Finally, the point C_v writes this information in the NFC tag.

  (c_4) *Updating the database*: This is a simple operation allows notifying to the server that a particular tag has crossed a control point. This makes possible a real-time tracking of the luggage.

According to the necessities of the cryptographic system, the use of a lightweight encryption with the addition homomorphic property is convenient. As shown in Figure 3, in each control point it is necessary to add a value to received message. This addition allows indicating only that that bag has gone across a new control point, but this change does not identify what point it is. This information is provided by the signature. For the guarantee of the signature verification, each control point must know the previous point in order to apply the verification operation with the correct keys. If the server needs decrypting the message for some reason, the result of this operation must be the addition of all the points where that bag has been. For this reason, the Paillier cryptosystem [8] has been chosen in order to fulfil this requirement.

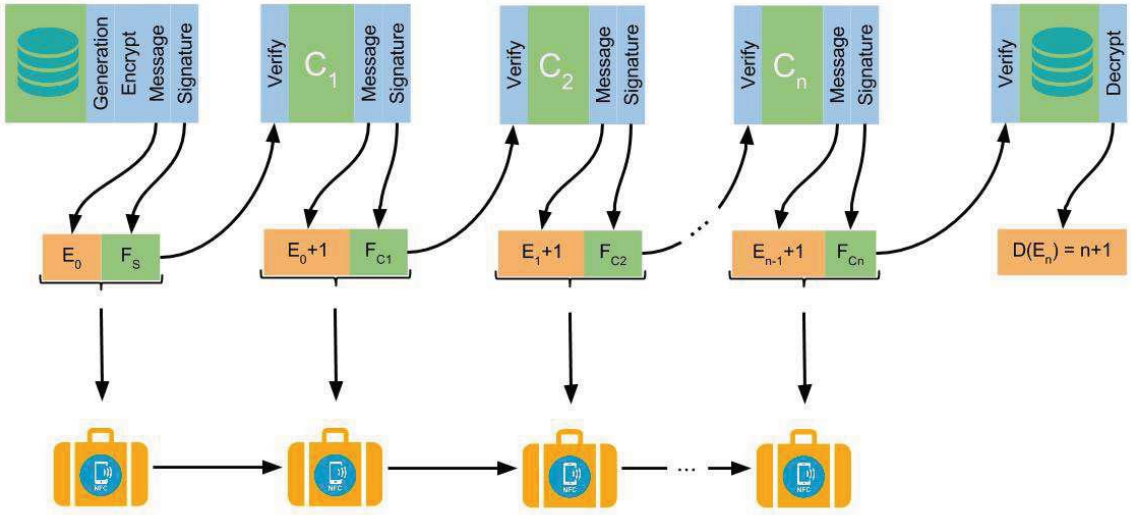


Figure 3: Verification and Update of Data in each Control Point

3.1 Paillier Cryptosystem

The Paillier cryptosystem is an asymetric scheme based on the problem of computing n-th residue classes. It is composed of three operations: key generation, encryption/decryption and digital signature/verify as follows:

Key generation

1. Choose two large primes p and q verifying that

$$gcd\Big(pq, (p-1)(q-1)\Big) = 1$$

2. Compute
$$n = pq, \lambda = lcm(p-1, q-1) = \frac{(p-1)(q-1)}{gcd(p-1, q-1)}$$
3. Select a random generator g where $g \in \mathbb{Z}^*_{n^2}$ verifying that n divides the order of g. This can be checked through the existence of the modular multiplicative inverse as follows:
$$\mu = \Big(L(g^\lambda \ mod \ n^2)\Big)^{-1} \ mod \ n$$
where the function L is defined as
$$L(u) = \frac{u-1}{n}$$
4. The generated key pair is composed of the public key $U = (n, g)$ and the private key $R = (\lambda, \mu)$.

Encryption

$$C_i = g^{M_i} \cdot r^n \ mod \ n^2$$

where

$$r \in \mathbb{Z}^*_n$$

Decryption

$$M_i = L(C_i^\lambda \ mod \ n^2) \cdot \mu \ mod \ n$$

The most interesting properties of Paillier cryptosystem for the present application are its homomorphic properties because, in particular, the addition of plaintexts is fundamental for this system, and the Paillier scheme verifies that

$$E(M_1) \cdot E(M_2) = (g^{M_1} \cdot r_1^n) \cdot (g^{M_2} \cdot r_2^n) = E(M_1 + M_2 \ mod \ n),$$

such that in the present application it can be extended as follows where P is the total number of control points:

$$\prod_{i=1}^{P} E(M_i) = \prod_{i=1}^{P} g^{M_i} \cdot r_i^n = E\Big((\sum_{i=1}^{P} M_i) \ mod \ n\Big)$$

In particular, for the proposal, this property involves that

$$D\Big(E(M_i) \cdot g^P \ mod \ n^2\Big) = M_i + P \ mod \ n$$

In this way, after the decryption in the last delivery control point, it can be checked that the luggage has gone through all the control points.

Digital signature

$$\begin{cases} t_1 = \dfrac{L\left(H(M_i)^{\lambda} \ mod \ n^2\right)}{L\left(g^{\lambda} \ mod \ n^2\right)} \quad mod \quad n \\ t_2 = \left(H(M_i) \cdot g^{-t_1}\right)^{\frac{1}{n \ mod \ \lambda}} \ mod \ n \end{cases}$$

where the pair (t_1, t_2) composes the signature. $H(M_i)$ is the hash function over the message which must fulfill that $H(M_i) \in \mathbb{Z}^*_{n^2}$.

Verification

$$H(M_i) \stackrel{?}{=} g^{t_1} \cdot t_2^n \ mod \ n^2$$

A didactic example is now included using small parameters in order to illustrate Paillier cryptosystem. In the key generation, we use $p = 7, q = 11$ and compute $n = p \cdot q = 77, \lambda = lcm(p-1, q-1) = 30$. Next, a g generator must be chosen so that $g \in \mathbb{Z}^*_{n^2}$, so we choose $g = 5652$ and compute $\mu = (L(g^{\lambda} \ mod \ n^2))^{-1} \ mod \ n = 74$.

Now, the key pair is composed by the public key which is $pkU = (n, g) = (77, 5652)$ and the private key which is $pkR = (\lambda, \mu) = (30, 74)$. With the key pair generated, we can cipher and decipher any value and also we can apply the homomorphic properties as follows. If we need to cipher $m = 23$, we must generate a random r value belonging to $\mathbb{Z}^*_n$, for example $r = 23$, and after we need to compute

$$c_1 = g^{M_i} \cdot r^n \ mod \ n^2 = 4304$$

Then, we can apply the homomorphic addition property with two values $a = 2$ and $b = 9$ and the decryption of the result is the addition of three values:

$$c_2 = c_1 \cdot g^a \ mod \ n^2 = 2245,$$

$$c_3 = c_2 \cdot g^b \ mod \ n^2 = 3404,$$

$$m = L(c_3^{\lambda} \ mod \ n^2) \cdot \mu \ mod \ n = 34$$

4 Vulnerabilities Analysis

Some possible vulnerabilities and the way that they have been taken into account both in the technological environment and in the system itself are exposed in this section.

In relation to the problems with the technology itself, two problems have been detected that do no suppose a great difference in comparison with the current systems with printed bar codes labels. The first one is related to the physical protection and, this aim is out of the scope of this system. If a passenger goes to the delivery point and his or her bag has not the NFC tag put correctly, this bag will be considered in an equivalent way as if in the present system a printed bar code is not available. For that, in this work it has been considered that passengers will have their belongings controlled in any moment. Besides, if a passenger comes to a delivery point without his or her NFC tag in the bag, the more convenient procedure would be applied according to each installation, for example using the current printed labels method. The second detected problem is the possibility to demagnetize NFC tag or put it in an illegible or corrupt state. In this case, the first control point verifies that the information stored in the tag is correct in the moment that the passenger puts his or her bag in the delivery point. If the verification was negative due to a impossibility to read the data or an incorrect signature verification, the responsible person would act using as in previous case. Also, coexistence of this system together with the present one could be a support system.

Some attacks could be launched on the technology itself. Thus, the possibility that malicious users try to change the content of foreign NFC tag exists. A huge number of mobile phone applications that allow managing this technology does this attack possible. In order to protect the system against these attacks, the digital signature of the message can be used to detect this situation when the stored information changes improperly.

Throughout the process, the data stored in NFC tag as well as its signature are modified in each control point. This provides a mechanism for controlling the baggage. Control of staff is not the objective of this work, the only aim of this work is to provide security in the management baggage. Consequently, the system guarantee the security against external attacks and assumes the reliability of personnel who manages luggage, in the same way as currently occurs.

5 Conclusions

This work describes a proposal for the permanent labelling of luggage that provides the ability to add new information along the control points for speeding up the passengers' flow and at the same time establishing more security over the control of luggage. The complete tracking of luggage from the origin to the destination is not contemplated in any existing system or previous proposals. This continuous tracking is here possible thanks to

the use of homomorphic functions. NFC technology is ised in the proposed system because it is becoming more accepted and economically affordable. This technology can be used to provide more control and more security over the luggage in the airports around the world. Besides, not only it provides more security against terrorist threats based on a low-control of luggage but also it is an answer to the high number of complaints for delays caused by luggage check-in and loss. Research on other NFC tags used to manage information access like MIFARE DESFire family, is an objective that keeps this work open.

Acknowledgements

Research supported by RTC-2014-1648-8, TEC2014-54110-R, MTM-2015-69138-REDT and DIG02-INSITU.

References

[1] L. Bi, Z. Feng, M. Liu, W. Wang: *Design and implementation of the airline luggage inspection system base on link structure of QR code*. IEEE International Symposium on Electronic Commerce and Security, pp. 527-530. (2008).

[2] Amadeus: *Reinventing the Airport Ecosystem. A new airline industry report'*. (2012).

[3] P. Seidenstat: *Terrorism, airport security, and the private sector*. Review of Policy Research, 21(3), pp. 275-291. (2004).

[4] W.K. Viscusi, R.J. Zeckhauser: *Sacrificing civil liberties to reduce terrorism risks*, pp. 1-22. (2003).

[5] G.E. Vastianos, D.M. Kyriazanos, V.I. Kountouriotis, S.C. Thomopoulos: *An RFID-based luggage and passenger tracking system for airport security control applications*. International Society for Optics and Photonics SPIE Defense+ Security, pp. 90911A-90911A. (2014).

[6] NTAG213/215/216: *NFC Forum Type 2 Tag compliant IC with 144/504/888 bytes user memory*.

[7] C.Gentry: *A fully homomorphic encryption scheme*. Doctoral dissertation, Stanford University. (2009).

[8] P. Paillier: *Public-key cryptosystems based on composite degree residuosity classes.* Advances in cryptology— EUROCRYPT'99, pp. 223-238. (1999).

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Natural grid numerical methods revisited in cell population balance models with asymmetric division

O. Angulo[1], J. C. López-Marcos[1] and M. A. López-Marcos[1]

[1] *Departamento de Matemática Aplicada, Universidad de Valladolid*

emails: oscar@mat.uva.es, lopezmar@mac.uva.es, malm@mac.uva.es

Abstract

We introduce and analyze a numerical method based on the integration along characteristics curves with the use of the *natural grid*. It is employed to obtain the solution to a cell population balance model structured by the cell size and with assymetric division rate.

Key words: population balance models, cell population dynamics, numerical methods, natural grid

1 Extended abstract

Cell population balance models were introduced in the early sixties and quickly experimented a great progress. The first efforts that can be associated with the concepts of population balance equations (PBE) were carried out within the framework of particle dynamics in chemical and cellular contexts [14, 4]. On the one hand, Randolph [14] formulates a generic population balance model (PBM) based on the generalized mechanical framework. This model is concerned about particles growth, aggregation and breakage, among others. On the other hand, Bell and Anderson [4] develop a cell population balance model in parallel. It is based on the cells growth and the probabilities of division and death of cells. In such a model, they consider two different state variables, cell-age and cell-volume, because of the different nature of reaccions, for instance involving DNA or ATP. In such a work, the division of an individual mother cell is considered to be into equally sized daughter cells. Later, Ramkrishna [12] adopts the generalized continuum mechanics concepts in [14] to derive a general PBE. He uses statistical concepts such as probability and expectancy to derive a set of equations in which age and cell mass are the descriptors of the cells state,

even though they may only indirectly be indicators of the cell metabolism, and reflect the influences of other biochemical substances. The division (or breakage birth) term employed in is more general than in [4]. From a formal point of view, cell PBMs typically consist of a PBE along with boundary and initial conditions as well as other coupled equations describing cell division probability and intensity, partitioning cell content upon division probability, stage transitions and, in the case of chemically structured models, cellular kinetics. PBE can be defined as the balance equation that accounts for the various processes that change the number of cells in a population and takes the form of first-order partial integro-differential equation, while supplementary equations, coupled in a non-linear way, are tipically ordinary integro-differential equations.

From a theoretical point of view, mathematical treatment of linear cell PBM has been developed since the early eighties. The study of the well-posedness of the size-structured problem with division into equally sized daughter cells and its convergence towards an asymptotically *stable-type size distribution* was made in [5]. The PBM model proposed by Ramkrishna was studied similarly in [6]. In the case of nonlinear models, where the vital rates, sources and sinks depends explicitly on the environment, the theoretical properties of existence and uniqueness of solutions are also needed [6]. A general survey of the main mathematical problems solved and the principal techniques employed in this context is given in [9, 8, 3, 10].

In spite of this early development, nowadays population balance modeling is an area of increasing aplications and is currently used to describe quite different issues. Ramkrishna [13] did investigate biological populations as well as numerous studies of chemical engineering problems had been reported. These problems demostrate the wide applicability of the PBM modeling framework and suggest that many potential applications remain unexplored.

The practical application of cell PBM is not easy due to the fact that owing to their complex mathematical nature (first-order partial integro-differential equations, sometimes coupled in a nonlinear way with ordinary integro-differential equations). The development of numerical algorithms for the accurate approximation of their solution is a challenging task. In the last twenty years, several studies have addressed their numerical solution with different techniques: analytical solutions based on a successive generations approach in the case of simple single-cell growth expresions, classical finite difference schemes, finite element or spectral methods or the use of the integration along the characteristics (see [1] and references there in). However, the analysis of most of these numerical proposals is not finished yet and a convergence theorem to the theoretical solution has been provided only in a few of them [2]. This theoretical issue will lead to a variety of algorithms that may be used to efficiently obtain accurate solutions of these models and hence facilitate their use.

Here we consider a simple PBM where the evolution of a cell population is structured by only one state variable (the cell size) where cells reproduction is by fission into different

daughter cells [11]. Cell size is an attractive parameter due to the relative ease and precision with which it can be measured because the instrumentation for obtaining it has improved considerably. The model that arises due to this simplification is still useful in order to analyze and understand the cell population dynamics.

We consider a nonnegative minimum cell size x_m and a maximal size, normalized to 1, at which point every cell might divide or die. We suppose the cell does not divide until it reaches a minimal size a, so $0 \leq x_m \leq a < 1$. We also assume that the environment is unlimited and all possible nonlinear mechanisms are ignored. The problem is given by the PBE

$$u_t(x,t) + (g(x)\,u(x,t))_x = -\mu(x)\,u(x,t) - b(x)\,u(x,t) + 2\int_x^1 b(s)\,P(x,s)u(s,t)\,ds, \quad (1)$$

$x_m < x < 1$, $t > 0$, a boundary condition

$$u\,(x_m,t) = 0, \quad t > 0, \quad (2)$$

and an initial size distribution

$$u(x,0) = \varphi(x), \quad x_m \leq x \leq 1. \quad (3)$$

The independent variables x and t represent size and time, respectively. The dependent variable $u(x,t)$ is the size-specific density of cells with size x at time t and we assume that the size of any individual varies according to the following ordinary differential equation

$$\frac{dx}{dt} = g(x). \quad (4)$$

The nonnegative functions g, μ and b represent the growth, mortality and division rate, respectively. These are usually called the vital functions and define the life history of the individuals. Note that all of them depend on the size x (the internal structuring variable). The dispersion of sizes at division amongst the two daughter cells (unequal division) is defined in terms of the partitioning function $P(x,y)$, a probability density function which gives the distribution of the size of a daughter-cell x when the size of the mother is equal to y. It satisfies the following conditions:

$$\int_{x_m}^1 P(x,y)\,dx = 1, \qquad P(x,y) = P(y-x,y), P(x,y) = 0, \quad x \geq y. \quad (5)$$

In accordance with accepted biological point of view, there exists a maximum size. This means that cells will divide or die with probability one before reaching it. To this end, if μ and b are positive and bounded functions, we consider a growth function, introduced by Von Bertalanffy, satisfying $\lim_{x\to 1}\int_{x_m}^x \frac{ds}{g(s)} = +\infty$. Note that if g is a continuous function

 ISBN: 978-84-608-6082-2

defined in $[x_m, 1]$ then this hypothesis implies that $g(1) = 0$. Moreover, the solution to the problem must satisfy $u(1, t) = 0$, $t > 0$, because we suppose that initially there are no cells of maximum size.

We will propose and analyze a characteristic curves scheme which employs a suitable invariant, nonuniform grid on the space variable, usually called *the natural grid*, as it was considered in [7]. This grid is quite interesting because its invariance allows us to study, at least experimentally, the long time behaviour of the cell population. We will provide numerical experiments which confirm the predicted accuracy of the numerical scheme.

Acknowledgements

This work was supported in part by projects MTM2014-56022-C2-2-P of the Ministerio de Economía y Competitividad (Spain) and European FEDER Funds and VA191U13 of the Junta de Castilla y León (Spain).

References

[1] O. Angulo, J. C. López-Marcos, M. A. López-Marcos, *A semi-Lagrangian method for a cell population model in a dynamical environment*, Math. Comput. Model. **57** (2013) 1860–1866.

[2] O. Angulo, J. C. López-Marcos, M. A. López-Marcos, *A second-order method for the numerical integration of a size-structured cell population model*, Abstr. Appl. Anal. **549168** (2015) 1–8.

[3] O. Arino, *A survey of structured cell population dynamics.* Acta Biotheor. **43** (1995) 3–25.

[4] G. I. Bell, E. C. Anderson, *Cell growth and division: I. a mathematical model with applications to cell volume distributions in mammalian suspension cultures*, Biophys. J. **7** (1967) 329–351.

[5] O. Diekmann, H. J. A. M. Heijmans, H. R. Thieme, *On the stability of the cell size distribution.* J. Math. Biol. **19** (1984) 227–248.

[6] H. J. A. M. Heijmans, *On the Stable Size Distribution of Populations Reproducing by Fission into Two Unequal Parts.* Math. Biosci. **12** (1984) 119-150.

[7] K. Ito, F. Kappel, G. Peichl, *A fully discretized approximation scheme for size-structured population models.*, SIAM J. Numer. Anal. **28** (1991) 923-954.

[8] A. Lasota, M. C. Mackey, *Probabilistic Properties of Deterministic Systems.* Cambridge University Press, London, 1985.

[9] J. A. J. Metz, O. Diekmann (Eds.), *The Dynamics of Physiologically Structured Populations.* Lect. Notes Biomath. **68**, Springer-Verlag, New York, 1986.

[10] B. Perthame, *Transport Equations in Biology.* Birkhuser, Basel, Switzerland, 2007.

[11] D. Ramkrishna, *Statistical models of cell populations.* In: *Adv. Biochem. Eng.* **11** 1–47, Springer, Berlin, 1979.

[12] D. Ramkrishna, *Population balances: theory and applications to particulate systems in engineering.* Academic Press, San Diego, 2000.

[13] D. Ramkrishna, M. R. Singh, *Population Balance Modeling: Current Status and Future Prospects* Annu. Rev. Chem. Biomol. Eng. **5** (2014) 123-146.

[14] A.D. Randolph, *A population balance for countable entities.* Can. J. Chem. Eng. **42** (1964) 280–281.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Distribution function estimates from dual frame context

Antonio Arcos[1], Sergio Martínez[2], María del Mar Rueda[1] and Helena Martínez[2]

[1] *Department of Statistics and Operational Research, University of Granada, Spain*

[2] *Department of Mathematics, University of Almería, Spain*

emails: arcos@ugr.es, spuertas@ual.es, mrueda@ugr.es, hmp603@ual.es

Abstract

The estimation of a finite population distribution function under a dual frame context is considered when auxiliary population information is available. Several procedures are defined and compared to a numbers of methods which have been adapted from the literature. The asymptotic distribution of proposed estimators is established and a brief simulation study is also included.

Key words: Auxiliary information, calibration technique, distribution function estimates, dual frames surveys

1 Introduction

The focus of this paper is on the estimation of the finite population distribution function, on the basis of a sample taken from a dual frame context. The subject is important: the distribution function is a basic statistic underlying many others ([1]); for purposes of assessing and comparing finite populations it can be more revealing than means and totals ([2]).

In words, for a given value of t, the finite distribution function $F(t)$ is the proportion of y in the population of size N not exceeding t (usually defined using the indicator function $\Delta(u)$ defined by $\Delta(u) = 1$, if u is true, $\Delta(u) = 0$, if not. The basic properties estimating the finite distribution function $F(t)$ are: it is monotonic nondecreasing in t, it is a step function with step size $1/N$ and its values are confined to $[0, 1]$. Estimating the finite population distribution function $F(t)$ is in some respects easier and in others more difficult than estimating a population total or mean. On the one hand, for fixed t, $F(t)$ is simply a

 ISBN: 978-84-608-6082-2

mean of 0's and 1's. On the other hand, we typically want to estimate $F(t)$ for more than one value of t and these estimates need to be coordinated and where the main variable y is related to an auxiliary variable x, it becomes an issue how to use this information, since now we are concerned with $\Delta(y \leq t)$ and not y itself, which is what usually gets modeled on x. Usually, the estimators of $F(t)$ are constructed with reference to some specific linear regression model with "slope" β which is estimated by weighted least squares using design weights. With the fitted values from the regression available, model calibrated estimators can be defined.

In a dual frame context there are several ways to obtain design weights and there are also several situations to relate the auxiliary information from frames. Now the main variable y is related to an auxiliary variable x in each frame, and this auxiliary variable can be different from frames or it can be the same. There has been no comprehensive comparison of the many available alternatives for estimating $F(t)$.

In this paper we adapt distribution function estimates in dual frame context with no auxiliary information, with the auxiliary information about the frame sizes and with more auxiliary information. In this last case, a post-stratification estimator with partial auxiliary information (as in [3]) and a model calibration estimator with complete auxiliary information (following [4] and [5]), are defined. The asymptotic distribution of proposed estimators is established. A brief simulation study is included in section 3.

2 Distribution function estimates in dual frame context

Assume we have a finite set of N population units identified by integers, $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$, and let A and B be two sampling-frames, both can be incomplete, but it is assumed that together they cover the entire finite population. Let $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ the set of population units in frame B. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$, where ^{c} denotes the complementary of a set. Let N, N_A, N_B, N_a, N_b and N_{ab} be the number of population units in $\mathcal{U}, \mathcal{A}, \mathcal{B}, a, b$ and ab, respectively.

Let y be a variable of interest in the population and let y_k be its value on unit k, for $k = 1, ..., N$. The objective is to estimate the finite population distribution function

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \quad \text{where} \quad \Delta(t - y_k) = \begin{cases} 1 & \text{if } t \geq y_k \\ 0 & \text{if } t < y_k \end{cases} \tag{1}$$

that can be written as the population mean of Δ_t values. Let D_{t_k} be the Δ_t value on unit k, for $k = 1, ..., N$. Then, the objective is to estimate, for each t, the population mean, $\overline{D}_t = D_t/N$, where D_t denotes the population total of variable D_t. When the population size N is unknown, N can be viewed as the total of constant $\mathbf{1}_N$ or $N = D_\infty$. This reduce our goal to estimate the population totals D_t.

 ISBN: 978-84-608-6082-2

This population total D_t can be written as $D_t = D_{t_a} + D_{t_{ab}} + D_{t_b}$, where $D_{t_a} = \sum_{k \in a} D_{t_k}, D_{t_{ab}} = \sum_{k \in ab} D_{t_k}$ and $D_{t_b} = \sum_{k \in b} D_{t_k}$. To this end, independent samples s_A and s_B are drawn from frame A and frame B of sizes n_A and n_B, respectively. Unit k in $\mathcal{A}$ has first-order inclusion probability $\pi_k^A = Pr(k \in s_A)$ and unit k in $\mathcal{B}$ has first-order inclusion probability $\pi_k^B = Pr(k \in s_B)$.

From data collected in s_A, it is possible to compute one unbiased estimator of the total for each domain in frame A, $\hat{D}_{t_a}$ and $\hat{D}_{t_{ab}}^A$, as described below:

$$\hat{D}_{t_a} = \sum_{k \in s_A} \delta_k(a) d_k^A D_{t_k}, \qquad \hat{D}_{t_{ab}}^A = \sum_{k \in s_A} \delta_k(ab) d_k^A D_{t_k},$$

where $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise and d_k^A are the weights under the sampling design used in frame A, defined as the inverse of the first order inclusion probabilities, $d_k^A = 1/\pi_k^A$. Similarly, using information included in s_B, one can obtain an unbiased estimator of total for domain b and another one for domain ab, $\hat{D}_{t_b}$ and $\hat{D}_{t_{ab}}^B$, which can be expressed as

$$\hat{D}_{t_b} = \sum_{k \in s_B} \delta_k(b) d_k^B D_{t_k}, \qquad \hat{D}_{t_{ab}}^B = \sum_{k \in s_B} \delta_k(ab) d_k^B D_{t_k},$$

with $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise, and d_k^B the weights under the sampling design used in frame B defined as the inverse of the first order inclusion probabilities, $d_k^B = 1/\pi_k^B$.

Different approaches for estimating the population total from dual frame surveys have been proposed in the literature. Below, we adapt some of them to the context of the distribution function estimation.

2.1 No auxiliary information

Dual-frame approach

The dual frame approach [6], [7]) suggests a convex combination of the two overlap estimates to obtain an unbiased global estimator of the population total.

Using this idea we consider the use of a fixed quantity η to weight $\hat{D}_{t_{ab}}^A$ and $\hat{D}_{t_{ab}}^B$, providing the estimator

$$\hat{D}_{t_\eta} = \hat{D}_{t_a} + (\eta)\hat{D}_{t_{ab}}^A + (1-\eta)\hat{D}_{t_{ab}}^B + \hat{D}_{t_b}. \tag{2}$$

This approach can always be applied since it does not require any previous or additional information but only the choice of the parameter η. Different procedures for selecting the value of η can be considered, yielding to different estimators. A simple selection is $\eta = 1/2$ that yields to the known as Fixed Weight Adjustment (FWA) estimator.

Single-frame approach

[8] and [9] combine all sampling units coming from the two frames, s_A and s_B, trying to build a single sample as if it was drawn from only one frame. Sampling weights for the units in the overlap domain need, then, to be modified to avoid bias. These adjusted weights are

$$\tilde{d}_k = \begin{cases} d_k^A & \text{if } k \in a \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in ab \\ d_k^B & \text{if } k \in b \end{cases}. \tag{3}$$

Hence, a new estimator of the population total can be defined in the form

$$\hat{D}_{t_{BKA}} = \sum_{k \in s_A} \tilde{d}_k^A D_{t_k} + \sum_{k \in s_B} \tilde{d}_k^B D_{t_k} = \sum_{k \in s} \tilde{d}_k D_{t_k}, \tag{4}$$

with $s = s_A \cup s_B$. Note that to compute this estimator, one needs to know, for units in sample coming from the overlap domain, the inclusion probability under both sampling designs.

2.2 Estimation with the auxiliary information about the frame sizes

The PML method for the dual-frame approach

[10] use a pseudo maximum likelihood approach to extend to complex designs the maximum likelihood estimator proposed by [6] only for simple random sampling without replacement. We adapt this estimator to the distribution function, and the required estimator of the population total is given by

$$\begin{aligned} \hat{D}_{t_{PML}} = & \frac{N_A - \hat{N}_{ab}^{PML}(\gamma)}{\hat{N}_a^A} \hat{D}_{t_a}^A + \frac{N_B - \hat{N}_{ab}^{PML}(\gamma)}{\hat{N}_b^B} \hat{D}_{t_b}^B \\ & + \frac{\hat{N}_{ab}^{PML}(\gamma)}{\gamma \hat{N}_{ab}^A + (1-\gamma)\hat{N}_{ab}^B} [\gamma \hat{D}_{t_{ab}}^A + (1-\gamma)\hat{D}_{t_{ab}}^B], \end{aligned} \tag{5}$$

where $\hat{N}_{ab}^{PML}(\gamma)$ is the smallest of the roots of quadratic equation $[\gamma/N_B + (1-\gamma)/N_A]x^2 - [1 + \gamma \hat{N}_{ab}^A/N_B + (1-\gamma)\hat{N}_{ab}^B/N_A]x + \gamma \hat{N}_{ab}^A + (1-\gamma)\hat{N}_{ab}^B = 0$ and $\gamma \in (0,1)$. It is also shown that the following value for γ

$$\gamma_{opt} = \frac{\hat{N}_a N_B V(\hat{N}_{ab}^B)}{\hat{N}_a N_B V(\hat{N}_{ab}^B) + \hat{N}_b N_A V(\hat{N}_{ab}^A)} \tag{6}$$

minimizes the variance of $\hat{D}_{t_{PML}}$.

The calibration method for the dual-frame approach

Recently, [11] extended calibration procedures to estimation from dual frame sampling assuming that the population frame sizes (N_A, N_B) or the population domain sizes (N_a, N_{ab}, N_b) are known. For example, suppose that the dimension of the three sets N_A, N_B and N_{ab} is known. Then, we can build the auxiliary vector using domain membership indicator variables, i.e.

$$\mathbf{d}_k = \big(\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b)\big), \quad \text{for} \quad k = 1, \ldots, N, \tag{7}$$

where $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise, $\delta_k(ba) = 1$ if $k \in ba$ and 0 otherwise and $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise. Note that because the population units in the overlap domain ab can be sampled in either or both surveys, it is useful to create a duplicate domain $ba = \mathcal{B} \cap \mathcal{A}$, which is identical to $ab = \mathcal{A} \cap \mathcal{B}$, to denote the domain in the overlapping area from frame B. Thus, $N_{ba} = N_{ab}$.

To obtain the final weights that can be used directly to estimate population totals as in equation (9), let $\mathbf{D} = (N_a, \eta N_{ab}, (1-\eta)N_{ba}, N_b)$ be the vector of known totals. In this case, the calibration constraints are given by

$$\sum_{k \in s_a} w_k^{df} = N_a, \quad \sum_{k \in s_{ab}} w_k^{df} = \eta N_{ab}, \quad \sum_{k \in s_{ba}} w_k^{df} = (1-\eta) N_{ba}, \quad \sum_{k \in s_b} w_k^{df} = N_b. \tag{8}$$

In this context, the dual frame calibration estimator for the total required for distribution function estimation can be defined as follows

$$\hat{D}_{t_{df}} = \sum_{k \in s} d_k^{df} D_{t_k}, \tag{9}$$

where weights d_k^{df} are such that $min \sum_{k \in s} G(d_k^{df}, \breve{d}_k)$ subject to $\sum_{k \in s} d_k^{df} \mathbf{d}_k = \mathbf{D}$, with $G(\cdot, \cdot)$ a determined distance measure and

$$\breve{d}_k = \begin{cases} d_k^A & \text{if } k \in a \\ \eta d_k^A & \text{if } k \in ab \bigcap s_A \\ (1-\eta) d_k^B & \text{if } k \in ab \bigcap s_B \\ d_k^B & \text{if } k \in b \end{cases}, \tag{10}$$

being $\eta \in [0, 1]$.

The calibration method for the single-frame approach

Then, with a similar approach to that of $\hat{D}_{t_{BKA}}$, another calibration estimator can be computed as

$$\hat{D}_{t_{sf}} = \sum_{k \in s} d_k^{sf} D_{t_k}, \tag{11}$$

with weights d_k^{sf} verifying that $min \sum_{k\in s} G(d_k^{sf}, \tilde{d}_k)$ subject to $\sum_{k\in s} d_k^{sf}\mathbf{d}_k = \mathbf{D}$, being $\tilde{d}_k$ the weights defined in (3).

An estimator of the variance of any calibration estimator can be obtained through following expression

$$\hat{V}(\hat{D}_t) = \sum_{k\in s}\sum_{\ell\in s}((\pi_{k\ell} - \pi_k\pi_\ell)/\pi_{k\ell})(d_k^\star e_k)(d_\ell^\star e_\ell) \tag{12}$$

where $d_k^\star$ is given by (3) or by (10) according to whether we use $\hat{D}_{tsf}$ or $\hat{D}_{tdf}$, respectively.

In all previous cases and for each t, a point estimate, $\hat{F}_t$, and a variance estimate, $\hat{V}(\hat{F}_t)$ are obtained. Then, a confidence interval based on the pivotal method can be computed as $\hat{F}_t \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{F}_t)}$ where $z_{\alpha/2}$ is the critical value of a standard normal distribution.

2.3 Estimation with more auxiliary information

Poststratification

We consider here the problem of estimating $F(t)$ given sample values of main variable y together with auxiliary population information, x. Following [3], we consider a poststratified estimator with poststrata defined by the intervals of x. Let the L poststrata partitioning the population U be denoted $U_1, \cdots, U_L$, where $k \in U_l$ if $x_{(l-1)} < x_k \leq x_{(l)}$, and where $x_{(0)} = -\infty$, $x_{(1)} < x_{(2)} < \cdots < x_{(L-1)}$ are specified values and $x_{(L)} = \infty$. Let $s_1^a, \ldots, s_L^a$ be the corresponding partition of s_a so that $s_l^a = s_a \cap U_l$. Let N_l^a be the size of $U_l^a = U_l \cap a$ and let $\hat{N}_l^a = \sum_{k\in s_l^a} d_k^A$, $l = 1, \ldots, L$. Following a similar notation for domains ab, ba and b, a poststratified estimator of $F(t)$ under the dual frame approach is $\hat{F}_t = \hat{D}_{t_{PS}}/N$ with

$$\hat{D}_{t_{PS}}^{df} = \hat{D}_{ta}^{PS} + \eta\hat{D}_{tab}^{PS} + (1-\eta)\hat{D}_{tba}^{PS} + \hat{D}_{tb}^{PS} \tag{13}$$

where

$$\hat{D}_{ta}^{PS} = \sum_{l=1}^{L} \frac{N_l^a}{\hat{N}_l^a}\sum_{k\in s_l^a} d_k^A\Delta(t-y_k), \quad \hat{D}_{tab}^{PS} = \sum_{l=1}^{L} \frac{N_l^{ab}}{\hat{N}_l^{ab}}\sum_{k\in s_l^{ab}} d_k^A\Delta(t-y_k),$$

$$\hat{D}_{tba}^{PS} = \sum_{l=1}^{L} \frac{N_l^{ba}}{\hat{N}_l^{ba}}\sum_{k\in s_l^{ba}} d_k^B\Delta(t-y_k), \quad \hat{D}_{tb}^{PS} = \sum_{l=1}^{L} \frac{N_l^{b}}{\hat{N}_l^{b}}\sum_{k\in s_l^{b}} d_k^B\Delta(t-y_k).$$

Under a single-frame approach, in a similar way, we can define

$$\hat{D}_{t_{PS}}^{sf} = \sum_{l=1}^{L} \frac{N_l}{\hat{N}_l^{sf}}\sum_{k\in s} d_k^{sf}\Delta(t-y_k) \tag{14}$$

where $\hat{N}_l^{sf} = \sum_{k\in s_l} d_k^{sf}$.

Calibration

The auxiliary information provided by an auxiliary vector x can be used following [4] and [12], in the definition of calibration conditions. We consider a pseudo-variable $g_k = \widehat{\beta}' x_k$ for $k = 1, 2, \ldots, N$ where: $\widehat{\beta}' = (\sum_{k\in s} d_k q_k x_k x_k')^{-1} \cdot \sum_{k\in s} d_k q_k x_k y_k$ and q_k positive values. With the pseudo-variable g, we add in the minimization process the following conditions:

$$\sum_{k\in s} \omega_k \Delta(\mathbf{t}_g - g_k) = \sum_{k\in U} \omega_k \Delta(\mathbf{t}_g - g_k) \tag{15}$$

where $\mathbf{t}_g = (t_1, \ldots, t_P)'$ is a vector chosen arbitrarily assuming that $t_1 < t_2 < \ldots < t_P$.

For example, under the dual-frame approach, if we suppose that N_A, N_B and N_{ab} are known and an auxiliary vector x^A is available for units in frame A, we can compute the pseudo variable $g_k^A = \widehat{\beta}_A' x_k^A$ and similarly for g_k^B. Then, if we consider $\delta_k(A) = \delta_k(a) + \delta_k(ab)$ and $\delta_k(B) = \delta_k(b) + \delta_k(ba)$, the auxiliary vector (7), for $k = 1, \ldots, N$, can be extend to

$$\mathbf{dx}_k' = \big(\delta_k(A) \cdot \Delta(\mathbf{t}_g^A - g_k^A)', \delta_k(a), \delta_k(ab), \delta_k(B) \cdot \Delta(\mathbf{t}_g^B - g_k^B)', \delta_k(b), \delta_k(ba)\big), \tag{16}$$

with the conditions in (8) and the additional conditions as follows:

$$\sum_{k\in s_A} \omega_k \Delta(\mathbf{t_g^A} - g_k^A) = \mathbf{D}_g^A, \sum_{k\in s_B} \omega_k \Delta(\mathbf{t_g^B} - g_k^B) = \mathbf{D}_g^B$$

where $\mathbf{D}_g^A = \sum_{k\in U_A} \Delta(\mathbf{t}_g^A - g_k^A)$ and $\mathbf{D}_g^B = \sum_{k\in U_B} \Delta(\mathbf{t}_g^B - g_k^B)$, that is, the vector of known total is

$$\mathbf{DX}' = ((\mathbf{D}_g^A)', N_a, \eta N_{ab}, (\mathbf{D}_g^B)', N_b, (1-\eta) N_{ba}) \tag{17}$$

In a similar way, if we suppose that only N_A and N_B are known, the auxiliary vector can be expressed as: $\mathbf{dx}_k' = \big(\delta_k(A) \cdot \Delta(\mathbf{t}_g^A - g_k^A)', \delta_k(A), \delta_k(B) \cdot \Delta(\mathbf{t}_g^B - g_k^B)', \delta_k(B)\big)$ or equivalently

$$\mathbf{dx}_k' = \big(\delta_k(A) \cdot \Delta(\mathbf{T}_g^A - g_k^A)', \delta_k(B) \cdot \Delta(\mathbf{T}_g^B - g_k^B)'\big) \tag{18}$$

where $\mathbf{T}_g^A = (\mathbf{t}_g^A, g_{max}^A)$ with $g_{max}^A = \max_{k\in U_A} g_k^A$ and similarly for $\mathbf{T}_g^B$. Thus, the calibration conditions are:

$$\sum_{k\in s_A} \omega_k \Delta(\mathbf{T_g^A} - g_k^A) = \mathbf{D}_g^A, \sum_{k\in s_B} \omega_k \Delta(\mathbf{T_g^B} - g_k^B) = \mathbf{D}_g^B$$

and the vector of known total is

$$\mathbf{DX}' = \big((\mathbf{D}_g^A)', (\mathbf{D}_g^B)'\big) \tag{19}$$

 ISBN: 978-84-608-6082-2

In this context, the dual frame calibration estimator can be defined as follows:

$$\hat{D}_{t_{dfx}} = \sum_{k\in s} w_k^{dfx} D_{t_k}, \tag{20}$$

where weights w_k^{dfx} are such that $min \sum_{k\in s} G(w_k^{dfx}, \breve{d}_k)$ subject to $\sum_{k\in s} w_k^{df} \mathbf{dx}_k = \mathbf{DX}$, with $G(\cdot,\cdot)$ a determined distance measure distance satisfying the usual conditions required in the calibration approach, $\breve{d}_k$ are defined in (10), $\mathbf{DX}$ is given by (19) or (17) and $\mathbf{dx}_k$ is the auxiliary vector given by (16) or (18) respectively .

Thus, the calibrated weights obtained are given by:

$$w_k^{dfx} = F(\boldsymbol{\lambda}_{df}\mathbf{dx}_k) \tag{21}$$

where $F(u) = g^{-1}(u)$ denotes the inverse function of $g(w,d) = \partial G(w,d)/\partial w$ and the vector of Lagrange multipliers $\boldsymbol{\lambda}_{df}$ is determined with

$$\phi_s(\boldsymbol{\lambda}_{df}) = \sum_{k\in s} \breve{d}_k[F(\boldsymbol{\lambda}_{df}\mathbf{dx}_k) - 1]\mathbf{dx}_k = \mathbf{DX} - \widehat{\mathbf{DX}}_\eta \tag{22}$$

where $\widehat{\mathbf{DX}}_\eta = \sum_{k\in s} \breve{d}_k \mathbf{dx}_k$. In general, numerical methods are needed to solve the minimization problem for different distance measures. If we consider the chi-square distance $G(w_k^{dfx}, \breve{d}_k) = \frac{(w_k^{dfx} - \breve{d}_k)^2}{2\breve{d}_k}$ in the calibration process, the weights are given by:

$$w_k^{dfx} = \breve{d}_k + \breve{d}_k \boldsymbol{\lambda}_{df}\mathbf{dx}_k \tag{23}$$

and the vector of Lagrange multipliers $\boldsymbol{\lambda}_{df}$ is given by: $\boldsymbol{\lambda}_{df} = \left(\mathbf{DX} - \widehat{\mathbf{DX}}_\eta\right)' \cdot T_{df}^{-1}$ with $T_{df} = \sum_{k\in s} \breve{d}_k \mathbf{dx}_k \mathbf{dx}_k'$ and the resulting calibration estimator $\hat{D}_{t_{dfx}}$ with weights (23) is:

$$\hat{D}_{t_{dfx}} = \sum_{k\in s} w_k^{dfx} D_{t_k} = \hat{D}_{t_\eta} + \left(\mathbf{DX} - \widehat{\mathbf{DX}}_\eta\right)' \cdot \widehat{B}_{df} \tag{24}$$

with

$$\widehat{B}_{df} = T_{df}^{-1} \cdot \sum_{k\in s} \breve{d}_k \mathbf{dx}_k D_{t_k} \tag{25}$$

Thus, following ([4]), we have established the next theorem.

Theorem 2.1 *The estimator $\hat{D}_{t_{dfx}}$ is asymptotically normal and asymptotically design unbiased and its asymptotic variance is given by*

$$V(\hat{D}_{t_{dfx}}) = V\left(\sum_{k\in s} \breve{d}_k E_k\right) \tag{26}$$

where $E_k = D_{t_k} - \boldsymbol{dx}_k' B_{DF}$ and $B_{DF} = T_{DF}^{-1} \cdot \sum_{k\in U} \boldsymbol{dx}_k D_{t_k}$ with $T_{DF} = \sum_{k\in U} \boldsymbol{dx}_k \boldsymbol{dx}_k'$

In a similar way, under the single-frame approach, if we suppose that N_A, N_B and N_{ab} are known, we can defined a calibration estimator as follows:

$$\hat{D}_{t_{sfx}} = \sum_{k\in s} w_k^{sfx} D_{t_k}, \tag{27}$$

where weights w_k^{sfx} are such that $min \sum_{k\in s} G(w_k^{sfx}, \tilde{d}_k)$ subject to $\sum_{k\in s} w_k^{sfx}\mathbf{dx}_k = \mathbf{DX}$, with $G(\cdot,\cdot)$ a determined distance measure, $\tilde{d}_k$ are defined in (3),

$$\mathbf{dx}'_k = (\delta_k(A)\cdot\Delta(\mathbf{t}_g^A - g_k^A)', \delta_k(a), \delta_k(ab), \delta_k(B)\cdot\Delta(\mathbf{t}_g^B - g_k^B), \delta_k(b)) \tag{28}$$

and

$$\mathbf{DX}' = \left((\mathbf{D}_g^A)', N_a, N_{ab}, (\mathbf{D}_g^B)', N_b\right) \tag{29}$$

If we suppose that only N_A and N_B are known, then the auxiliary vector and the vector of known totals are given by

$$\mathbf{dx}'_k = \left([\delta_k(A)]\cdot\Delta(\mathbf{T}_g^A - g_k^A)', [\delta_k(B)]\cdot\Delta(\mathbf{T}_g^B - g_k^B)\right) \tag{30}$$

and

$$\mathbf{DX}' = \left((\mathbf{D}_g^A)', (\mathbf{D}_g^B)'\right) \tag{31}$$

Thus, the solution for the chi-square distance is given by:

$$w_k^{sfx} = \tilde{d}_k + \tilde{d}_k \boldsymbol{\lambda}_{sf}\mathbf{dx}_k \tag{32}$$

with $\boldsymbol{\lambda}_{sf} = \left(\mathbf{DX} - \widehat{\mathbf{DX}}_{BKA}\right)'\cdot T_{sf}^{-1}$ where $\widehat{\mathbf{DX}}_{BKA} = \sum_{k\in s}\tilde{d}_k\mathbf{dx}_k$ and $T_{sf} = \sum_{k\in s}\tilde{d}_k\mathbf{dx}_k\mathbf{dx}'_k$. The resulting calibration estimator $\hat{D}_{t_{sfx}}$ with weights (32) is:

$$\hat{D}_{t_{sfx}} = \sum_{k\in s} w_k^{sfx} D_{t_k} = \hat{D}_{t_{BKA}} + \left(\mathbf{DX} - \widehat{\mathbf{DX}}_{BKA}\right)'\cdot\widehat{B}_{sf} \tag{33}$$

with

$$\widehat{B}_{sf} = T_{sf}^{-1}\cdot\sum_{k\in s}\tilde{d}_k\mathbf{dx}_k D_{t_k} \tag{34}$$

where $\mathbf{DX}$ is given by (29) or (31) and $\mathbf{dx}_k$ is given by (28) or (30), respectively.

Thus, in this single frame context, we have established the next theorem.

Theorem 2.2 *In the same way as for dual frame estimator, the estimator $\hat{D}_{t_{sfx}}$ is asymptotically normal and asymptotically design unbiased and its asymptotic variance is given by*

$$V(\hat{D}_{t_{sfx}}) = V\left(\sum_{k\in s}\tilde{d}_k E_k\right) \tag{35}$$

where $E_k = D_{t_k} - \boldsymbol{dx}'_k B_{SF}$ and $B_{SF} = T_{SF}^{-1}\cdot\sum_{k\in U}\boldsymbol{dx}_k D_{t_k}$ with $T_{SF} = \sum_{k\in U}\boldsymbol{dx}_k\boldsymbol{dx}'_k$.

 ISBN: 978-84-608-6082-2

The precision of $\hat{D}_{t_{dfx}}$ and $\hat{D}_{t_{sfx}}$ calibration estimators changes with the selection of $\mathbf{t}_g^A$ and $\mathbf{t}_g^B$. Using only one point in the calibration process, it is possible to calibrate on the same t at we need to estimate the finite population distribution function, $F(t)$. We denote this calibration estimator by Cal1. Following [12] as alternative estimation method, we select only one optimum point in the calibration process and we denote the resulting calibration estimator by CAL1. Both calibration estimators can be applied under the single-frame and the dual-frame approach.

In a similar way, it is also possible to use two o more points. For example, two points (t_4 and t_8) or two optimal points, three points (t_3, t_6 and t_9) or three optimal points, and eleven points ($t_k, k = 1, \ldots, 11$). In [4] it is suggested that a fairly small number of arbitrarily selected points t_j may suffice, as [13] point out. In the next section we present a subset of results obtained in simulations studies to compare efficiency of proposed estimators.

3 Simulation study

We conducted a simulation study to analyze the performance of the proposed estimators. Our simulations are programmed in `R` using the `Frames2` package ([14]) to build the estimators. In addition, we developed some new `R`-code to compute the proposed calibration estimators for dual frame context.

The simulated population has a dimension $N = 2,350$. The resulting sizes of the two frames are N_A=1,735 and N_B=1,191 and, consequently, the overlap domain size is N_{ab}=601. The units from frame A are then divided into six strata as follows: $N_{Ah} = (727, 375, 113, 186, 115, 219)$. Samples from frame A are selected using stratified simple random sampling. Samples from frame B are selected by means of SRSWOR sampling. Sample sizes for frame A and frame B, which correspond to the following number of units per stratum are $n_A = (15, 20, 15, 20, 15, 20) = 105$ and $n_B = 135$.

The simulation study consisted of selecting $R = 1,000$ samples of size $n = 240$. From each sample and for each estimator, estimates of the distribution function $F(t)$ were calculated for 11 different values of t, namely the quantiles t_q of the population distribution function such that $q = 1/12, \ldots, 11/12$, $F(t_q) = q$.

Aggregated measures of performance used to summarize the results for the various quantiles are the average absolute bias (AVAB) and the average root mean squared error (AVRMSE), given respectively by

$$\text{AVAB} = \frac{1}{11}\sum_q |\frac{1}{R}\sum_{s=1}^{R}(\hat{F}^s(t_q) - F(t_q))|, \quad \text{AVRMSE} = \frac{1}{11}\sum_q |\sqrt{\frac{1}{R}\sum_{s=1}^{R}(\hat{F}^s(t_q) - F(t_q))^2}|.$$

All calibration estimators were computed using N_a, N_{ab}, N_b as auxiliary information and η as in (6), and are denoted by Cal. In addition, X_A and X_B variables are used in

the calibration process and the auxiliary information about one point (the same at we need the estimation), two points (t_4, t_8), and three points (t_3, t_6, t_9) is added, and the resulting calibration estimators are denoted by Cal1, Cal2 and Cal3. In similar way, when one, two or three optimal points are used, we denote the resulting calibration estimator by CAL1, CAL2 and CAL3, respectively. Post-stratified estimator based on X_A and X_B and three equidistant post-strata are also computed for the two approach. Finally, BKA estimator, under the Single-Frame approach, and FWA and PML estimators, under the Dual-Frame approach, are computed. Table 1 shows our results.

Table 1: Average absolute bias (AVAB) and Average root mean squared error (ARMSE) for several estimator over 1,000 samples

Single-Frame			*Dual-Frame*		
	AVAB	AVRMSE		AVAB	AVRMSE
BKA	46	362	FWA	45	372
			PML	48	366
Cal	46	362	Cal	46	364
Pos	109	225	Pos	106	231
Cal1	38	301	Cal1	38	307
Cal2	33	280	Cal2	33	283
Cal3	38	259	Cal3	38	265
CAL1	16	219	CAL1	16	229
CAL2	26	212	CAL2	26	221
CAL3	9	203	CAL3	9	216

Values multiplied by 10,000

The average absolute biases are negligible in all cases, as expected from the theoretical results. In terms of ARMSE, other things being equal, single-frame estimators are more efficient than dual-frame estimators, due to the extra-information they incorporate in the estimation process. Given a particular optimal point in the calibration process, it makes difference in terms of efficiency and this, again, is in line with our theoretical findings.

Acknowledgements

This study was partially supported by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía, Spain) and by Ministerio de Economía y Competitividad (grant MTM2015-63609-R).

References

[1] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York. 1970.

[2] N. Sedransk, J. Sedransk, *Distinguishing among distributions using data from complex sample designs*, J. Amer. Stat. Assoc. **74** (1979) (368) 754760.

[3] P. Silva, C. Skinner, *Estimating distribution functions with auxiliary information using poststratification*, J. Official Statist. **11** (1995) 277–294.

[4] M. Rueda, S. Martínez, H. Martínez, A. Arcos, *Estimation of the distribution function with calibration methods*, J. Statist. Plann. Inference **137** (2007) 435–448.

[5] S. Martínez, M. Rueda, H. Martínez, A. Arcos, *Determining p optimum calibration points to construct calibration estimators of the distribution function*, Journal of Computational and Applied Mathematics **275** (2015) 281–293.

[6] H. O. Hartley, *Multiple Frame Surveys*, Proceedings of the American Statistical Association, Social Statistics Sections (1962) 203–206.

[7] W. A. Fuller, L. F. Burmeister, *Estimation for samples selected from two overlapping frames*, in: ASA Proceedings of the Social Statistics Sections, (1972) 245–249.

[8] M. D. Bankier, *Estimators based on several stratified samples with applications to multiple frame surveys*, J. Amer. Stat. Assoc. **81** (1986) 1074–1079.

[9] G. Kalton, D. W. Anderson, *Sampling rare populations*, Journal of the Royal Statistical Society **149** (1986) 65–82.

[10] C. J. Skinner, J. N. K. Rao, *Estimation in dual frame surveys with complex designs*, J. Amer. Stat. Assoc. **91** (443) (1996) 349–356.

[11] M. G. Ranalli, A. Arcos, M. Rueda, A. Teodoro, *Calibration estimators in dual frames surveys*, Statistical Methods and Aplications Online First (2015).

[12] S. Martínez, M. Rueda, A. Arcos, H. Martínez , *Optimum calibration points estimating distribution functions* , Journal of Computational and Applied Mathematics **233** (2010) 2265–2277.

[13] C. E. Särndal, *The calibration approach in survey theory and practice*, Survey Methodology **2** 99–119.

[14] A. Arcos, D. Molina, M. Rueda, M. G. Ranalli, *Frames2: A package for estimation in dual frame surveys*, The R Journal **7** (1) (2015) 52–72.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Convergence of Newton's method under Vertgeim conditions: new extensions using restricted convergence domains

I. K. Argyros[1], J. A. Ezquerro[2], M. A. Hernández-Verón[2] and Á. A. Magreñán[3]

[1] *Department of Mathematics Sciences, Cameron University, USA.*

[2] *Department of Mathematics and Computation, University of La Rioja, Spain.*

[3] *School of Engineering, International University of La Rioja, Spain.*

emails: iargyros@cameron.edu, jezquer@unirioja.es, mahernan@unirioja.es, alberto.magrenan@unir.net

Abstract

We present new sufficient convergence conditions for the semilocal convergence of Newton's method to a locally unique solution of a nonlinear equation in a Banach space. We use Hölder and center Hölder conditions, instead of just Hölder conditions, for the first derivative of the operator involved in the convergence study. The new convergence conditions are weaker than others required in earlier studies.

Key words: Newton's method, semilocal convergence, recurrent functions, Hölder Hölder condition, center Hölder condition.
MSC 2000: 47H99, 65H10, 65J15.

1 Introduction

In general, when we consider the calculation of a solution of a nonlinear equation $F(x) = 0$, it is difficult to find an exact solution because there are hardly procedures that allow us to find it. For this reason, we usually apply iterative methods to approximate the solution. The best known and most used iterative method in practice is Newton's method, whose algorithm is:

$$\begin{cases} x_0 \text{ given,} \\ x_n = x_{n-1} - [F'(x_{n-1})]^{-1} F(x_{n-1}), \quad n \in \mathbb{N}. \end{cases}$$

So, the idea is clear: starting from a good initial approximation x_0 to a solution x^* of the equation $F(x) = 0$, we construct a sequence $\{x_n\}$ such that $x^* = \lim_{n\to\infty} x_n$, so that it is needed to impose enough conditions for the sequence is convergent.

To give sufficient generality to our study, we consider that $F : \Omega \subseteq X \longrightarrow Y$ is a nonlinear operator defined on a nonempty open convex subset Ω of a Banach space X with values in a Banach space Y. Many problems from computational sciences, physics and other disciplines can be brought into a form similar to equation $F(x) = 0$ using mathematical modelling. So, the unknowns of this equation can be functions (difference, differential, and integral equations), vectors (systems of linear or nonlinear algebraic equations), or real/complex numbers (single algebraic equations with single unknowns).

The fact of considering our study in Banach spaces has the advantage that all sequences of Cauchy in Banach spaces are convergent. Therefore, in our study, it deals with proving that $\{x_n\}$ is a sequence of Cauchy. For this, we study the semilocal convergence of Newton's method, so that conditions on the starting point x_0 and on the operator involved F are required, along with a condition that guarantee that the sequence $\{x_n\}$ is of Cauchy from the conditions required to x_0 and F.

2 Semilocal convergence of Newton's method

The first semilocal convergence result for Newton's method in Banach spaces was given by Kantorovich [4] under the following conditions:

(C1) There exists $\Gamma_0 = [F'(x_0)]^{-1} \in \mathcal{L}(Y, X)$, for some $x_0 \in \Omega$, with $\|\Gamma_0\| \leq \beta$ and $\|\Gamma_0 F(x_0)\| \leq \eta$, where $\mathcal{L}(Y, X)$ is the set of bounded linear operators from Y to X.

(C2) $\|F''(x)\| \leq K$ for $x \in \Omega$.

(C3) $h = K\beta\eta \leq \frac{1}{2}$ and $B(x_0, t^*) \subset \Omega$, where $t^* = \frac{1-\sqrt{1-2h}}{h}\,\eta$.

A few years later, Ortega observes that the second derivative of the operator involved, F'', is not in the algorithm of Newton's method and that condition (C2) implies that F' is Lipschitz continuous in Ω. In addition, he presents in [6] a variant of the result given by Kantorovich where (C2) is replaced by a Lipschitz condition for F' in Ω; that is:

(C2b) $\|F'(x) - F'(y)\| \leq K\|x - y\|$ for $x, y \in \Omega$.

Later, other authors ([1, 3, 5, 7]) consider the following generalization of the last condition:

(C2c) $\|F'(x) - F'(y)\| \leq \ell\|x - y\|^p$ for $x, y \in \Omega$ and $p \in [0, 1]$,

which is known as a Hölder condition for F' in Ω. Obviously, if $p = 1$, then condition (C2c) is reduced to condition (C2b).

This generalization of condition (C2b) leads to a modification of condition (C3) given for the parameters appeared previously. So, for example, Hernández proves in [3] the semilocal convergence of Newton's method under conditions (C1), (C2c) and

(C3b) $a = \ell\beta\eta^p \leq z^*$, where z^* is the unique zero of the function

$$f(t) = (1+p)^p(1-t)^{1+p} - t^p$$

in the interval $(0, 1/2]$ and $B(x_0, r) \subset \Omega$, where $r = \frac{(1+p)(1-a)}{(1+p)-(2+p)a}\,\eta$.

In this work, by means of some modifications of condition (C2c), we try to weaken condition (C3b). Observe that if condition (C2c) is verified in a domain $\Omega_0 \subset \Omega$, the value of the constant ℓ is obviously lower and, therefore, condition (C3b) is more easily verifiable. Even if we consider this situation in the extreme case, $\Omega_0 = \{x_0\}$, that is:

$$\|F'(x) - F'(x_0)\| \leq \ell_0\|x - x_0\|^p \quad \text{with} \quad x \in \Omega \quad \text{and} \quad p \in [0, 1],$$

it is clear that $\ell_0 \leq \ell$. Note that the last condition is called: center Hölder condition for F' in Ω. As a consequence, we do a brief reminder of the main modifications obtained for the previous condition and, using recurrent functions [2], obtain a result which improves the previously ones obtained by other authors.

Acknowledgements

This work has been partially supported by the project MTM2014-52016-C2-1-P of Spanish Ministry of Economy and Competitiveness.

References

[1] I. K. Argyros, *Remarks on the convergence of Newton's method under Hölder continuity conditions*, Tamkang J. Math. **23** (1992) 269–277.

[2] I. K. Argyros and S. Hilout, *On the weakening of the convergence of Newton's method using recurrent functions*, J. Complexity **25** (2009) 530–543.

[3] M. A. Hernández, *The Newton method for operators with Hölder continuous first derivative*, J. Optim. Theory Appl. **109** (2001) 631–648.

[4] L. V. Kantorovich and G. P. Akilov, *Functional analysis*, Pergamon Press, Oxford, 1982.

[5] H. B. Keller, *Numerical methods for two-point boundary value problems*, Dover Pub., New York, 1992.

[6] J. M. Ortega, *The Newton-Kantorovich theorem*, Amer. Math. Monthly **75** (1968) 658–660.

[7] J. Rokne, *Newton's method under mild differentiability conditions with error analysis*, Numer. Math. **18** (1972) 401–412.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Improving the domain of parameters for Newton's method

Ioannis K. Argyros[1], Á. Alberto Magreñán[2] and Juan Antonio Sicilia[2]

[1] *Department of Mathematics Sciences, Cameron University*

[2] *Escuela de Ingeniería, Universidad Internacional de La Rioja (UNIR)*

emails: iargyros@cameron.edu, alberto.magrenan@unir.net,
juanantonio.sicilia@unir.net

Abstract

We present a new technique to improve the convergence domain for Newton's method both in the semilocal. It turns out that with the new technique the error bounds are tighter and the information on the location of the solution is at least as precise as in earlier studies

Key words: Newton's method, Banach space, majorizing sequence, Kantorovich hypothesis, local-semilocal convergence.
MSC 2000: 65D10, 65D99, 65G99, 65H10,47J05, 47J25, 90C30

1 Introduction

In this study we are concerned with the problem of approximating a locally unique solution x^* of equation

$$F(x) = 0, \tag{1.1}$$

where F is a Fréchet-differentiable operator defined on a convex subset D of a Banach space X with values in a Banach space Y.

Many problems from Applied Sciences including engineering can be solved by means of finding the solutions of equations in a form like (1.1) using Mathematical Modelling [1, 2, 6, 8]. For example, dynamic systems are mathematically modeled by difference or differential equations, and their solutions usually represent the states of the systems. Except in special cases, the solutions of these equations can be found in closed form. This is the main reason

why the most commonly used solution methods are usually iterative. The convergence analysis of iterative methods is usually divided into two categories: semilocal and local convergence analysis. The semilocal convergence matter is, based on the information around an initial point, to give criteria ensuring the convergence of iteration procedures. A very important problem in the study of iterative procedures is the convergence domain. In general the convergence domain is small. Therefore, it is important to enlarge the convergence domain without additional hypothesis. Another important problem is to find more precise error estimates on the distances $\|x_{n+1}-x_n\|$, $\|x_n-x^*\|$.

Newton's method defined for each $n=0,1,2,\ldots$ by

$$x_{n+1}=x_n-F'(x_n)^{-1}F(x_n) \tag{1.2}$$

where x_0 is an initial point,is undoubtedly the most popular method for generating a sequence $\{x_n\}$ approximating x^*. Let $(U(z,\varrho)$, $\bar{U}(z,\varrho)$ stand, respectively for the open and closed ball in X with center $z\in X$ and radius $\varrho>0$. The best known semilocal convergence result for Newton's method is the Newton-Kantorovich theorem [6] (see Theorem 2.1 that follows) which is based on the hypotheses (given in affine invariant form) by:

(H_1) There exists $x_0\in D$ such that $F'(x_0)^{-1}\in L(Y,X)$ and a parameter $\eta\geq 0$ such that

$$\|F'(x_0)^{-1}F(x_0)\|\leq\eta$$

(H_2) There exists a parameter $L>0$ such that for each $x,y\in D$

$$\|F'(x_0)^{-1}(F'(x)-F'(y))\|\leq L\|x-y\|.$$

and

(H_3) $\bar{U}(x_0,R)\subseteq D$ for some $R>0$.

The sufficient convergence condition of Newton's method is given by the famous for its simplicity and clarity Kantorovich hypothesis

$$h=2l\eta\leq 1. \tag{1.3}$$

There are simple examples in the literature to show that hypothesis (1.3) is not satisfied but Newton's method converges starting at x_0 (See Example 2.2). Moreover, the convergence domain of Newton's method depending on the parameters L and η is in general small. Therefore it is important to enlarge the convergence domain by using the same constants L and η. Argyros et all in a series of papers [4, 5] presented weaker sufficient convergence conditions for Newton's method by using more precise majorizing sequences than before [6, 7]. These conditions are

$$h_1 = 2A_1\eta \leq 1, \tag{1.4}$$

$$h_2 = 2A_2\eta \leq 1, \tag{1.5}$$

$$h_3 = 2A_3\eta \leq 1, \tag{1.6}$$

$$h_4 = 2A_4\eta \leq 1, \tag{1.7}$$

where

$$A_1 = \frac{L_0 + L}{2}, \quad A_2 = \frac{1}{8}\left(L + 4L_0 + \sqrt{L^2 + 8L_0L}\right),$$
$$A_3 = \frac{1}{8}\left(4L_0 + \sqrt{L_0L + 8L_0^2} + \sqrt{L_0L}\right), \quad A_2 = \frac{1}{\eta_0},$$

η_0 is the small positive root of a quadratic polynomial (see Theorem in [1, 2, 4, 5] or Theorem 2.7 that follows) and $L_0 > 0$, is the center Lipschitz constant such that

$$\|F'(x_0)^{-1}(F'(x) - F'(x_0))\| \leq L_0\|x - x_0\| \quad \text{for each } x \in D \tag{1.8}$$

whose existence is always implied by (H_2).

We have that

$$L_0 \leq L \tag{1.9}$$

holds in general and $\frac{L}{L_0}$ can be arbitrarily large [3]. Notice also (1.8) is not an additional to (H_2) hypothesis, since in practice the computation of parameter L involves the computation of L_0 as a special case. Notice that if $L_0 = L$ conditions (1.4)–(1.7) reduce to condition(1.3). However, if $L_0 < L$, then we have [4, 5]

$$h \leq 1 \Rightarrow h_1 \leq 1 \Rightarrow h_2 \leq 1 \Rightarrow h_3 \leq 1 \Rightarrow h_4 \leq 1, \tag{1.10}$$

$$\begin{array}{l} \frac{h_1}{h} \to \frac{1}{2}, \quad \frac{h_2}{h} \to \frac{1}{4}, \quad \frac{h_2}{h_1} \to \frac{1}{2}, \quad \frac{h_3}{h} \to 0, \\ \text{and} \\ \frac{h_3}{h_2} \to 0, \quad \frac{h_3}{h_1} \to 0, \quad \text{as } \frac{L_0}{L} \to 0. \end{array} \tag{1.11}$$

Estimates (1.11) show by how many times (at most) a condition is improving the previous one.

Notice also that the error bounds on the distances involved as well as the location on the solution x^* are also improved under these weaker conditions [1, 2, 4, 5]. In the present study, the main goal is to improve further conditions (1.3)–(1.7) by using smaller than L_0 and L parameters and by restricting the domain D. Similar ideas are used to improve error bounds and enlarge convergence radii in the local convergence case.

2 Convergence Analysis

We present the semilocal convergence analysis of Newton's method. Next, we state the following version of the Newton Kantorovich theorem [6].

Theorem 1 *Let $F : \mathbb{D} \subset X \to Y$ be a continuously Fréchet-differentiable operator. Suppose that* (1.3) *and conditions* (H_1)*–*(H_3) *hold, where*

$$R = \frac{1 - \sqrt{1-h}}{L}.$$

Then, the sequence $\{x_n\}$ *generated by Newton's method is well defined, remains in* $U(x_0, R)$ *for each* $n = 0, 1, 2, \ldots$ *and converges to a unique solution* $x^* \in \bar{U}(x_0, R)$ *of equation* (1.1).

Let us consider an academic example, where the Newton-Kantorovich hypothesis (1.3) is not satisfied.

Example 1 *Let* $X = Y = \mathbb{R}$, $x_0 = 1$, $D = U(1, 1-p)$ *for* $p \in (0, \frac{1}{2})$ *and define function* F *on* D *by*

$$F(x) = x^3 - p.$$

We have that $\eta = \frac{1-p}{3}$ *and* $L = 2(2-p)$. *Then, hypothesis* (1.3) *is not satisfied, since*

$$h_k = \frac{4}{3}(2-p)(1-p) > 1 \quad \text{for each } p \in (0, \frac{1}{2}).$$

Hence, there is no guarantee that sequence $\{x_n\}$*starting from* $x_0 = 1$ *converges to* $x^* = \sqrt[3]{p}$.

Next, we present a semilocal convergence result that extends the applicability of Theorem 1.

Theorem 2 *Let $F : \mathbb{D} \subset X \to Y$ be a continuously Fréchet-differentiable operator. Suppose that there exist $x_0 \in D$, $\eta \geq 0$, $\gamma \geq 1$, $L_\gamma > 0$ such that*

$$F'(x_0) \in L(Y, X),$$

$$\|F'(x_0)^{-1} F(x_0)\| \leq \eta,$$

$$D_\gamma = U(x_0, \gamma\eta) \subseteq D,$$

$$\|F'(x_0)^{-1}(F'(x) - F'(y))\| \leq L_\gamma \|x - y\| \quad \text{for each } x, y \in D_\gamma,$$

$$h_\gamma = 2L_\gamma \eta \leq 1$$

and

$$R_\gamma \leq \gamma\eta,$$

where

$$R_\gamma = \frac{1-\sqrt{1-2h_\gamma}}{L_\gamma}.$$

Then, the sequence $\{x_n\}$ *generated by Newton's method is well defined, remains in* $U(x_0, R_\gamma)$ *for each* $n = 0, 1, 2, \ldots$ *and converges to a unique solution* $x^* \in \bar{U}(x_0, R_\gamma)$ *of equation* (1.1).

Example 2 *Returning back to the Example 1 let* $p = 0.49$ *and* $\gamma = 1.9$. *Then, we have that* $\eta = 0.17$, $\gamma\eta = 0.323$ *and* $R_\gamma = 0.258202394 < 0.323 < 1 - p = 0.51$. *Hence, the hypotheses of Theorem 2 are satisfied.*

In this work, we apply the same idea in order to improve the existing results related to conditions (1.4)–(1.7).

Acknowledgements

This research was supported by Universidad Internacional de La Rioja (UNIR, http://www.unir.net), under the Plan Propio de Investigación, Desarrollo e Innovación 3 [2015–2017]. Research group: MOdelación Matemática Aplicada a la INgeniería(MOMAIN), by the the grant SENECA 19374/PI/14 and by Ministerio de Ciencia y Tecnología MTM2014-52016-C2-01-P

References

[1] Argyros, I.K., "Convergence and Application of Newton-type Iterations," Springer, 2008.

[2] Argyros,I. K., S. Hilout, Numerical methods in Nonlinear Analysis, World Scientific Publ. Comp. New Jersey, 2013.

[3] Argyros, I. K., Concerning the terra incognita between convergence regions of two Newton methods, Nonlinear Anal., 62 (2005), 179-194.

[4] Argyros, I.K., Hilout, S., Weaker conditions for the convergence of Newton's method, Journal of Complexity, AMS, 28 (2012), 364–387.

[5] Argyros, I.K., Hilout, S., On an improved convergence analysis of Newton's method, Applied Mahtematics adn Computation, 225 (2013), 372–386.

[6] Kantorovich, L.V., Akilov, G.P., Functional Analysis, Pergamon Press, Oxford, 1982.

[7] Ren, H., Argyros, I. K., Convergence radius of the modified Newton method for multiple zeros under Hlder continuous derivative. Applied Mathematics and Computation, 217(2) (2010), 612–621 .

[8] J.F. Traub, Iterative methods for the solution of equations, Prentice- Hall Series in Automatic Computation, Englewood Cliffs, N. J., 1964.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

FDM for Stochastic Partial Differential Equations

Allaberen Ashyralyev[1]

emails: aashyr@fatih.edu.tr

Abstract

In present paper, a survey of recent results on difference schemes for approximate solutions of stochastic partial differential equations. Results on the convergence of difference schemes for parabolic, hyperbolic and telegraph equations are presented. Finally, I will formulate some problems and future plans.

Key words: difference schemes, stochastic partial differential equation, convergence estimates
MSC 2000: 65C30 (65M06)

1 Introduction

Methods for numerically solving the initial and the boundary value problems for stochastic ordinary and partial differential equations have been studied and developed over the last three decade. An excellent survey of work in this area was given by Kloeden and Platen [1] - [2]. It is known that most problems in heat flow, fusion process, model financial instruments like options, bonds and interest rates and other areas which are involved with uncertainty lead to stochastic differential equation with parabolic type. These equations can be derived as models of indeterministic systems and consider as methods for solving boundary value problems [3]- [4].

The approach of the book [5] permitted essentially to extend a class of problems where the theory of difference methods is applicable. Namely, now it is possible to investigate the stochastic differential equations (see, for examples, [6]-[13]). In present paper, a survey of recent results on difference schemes for approximate solutions of stochastic partial differential equations are given. Results on the convergence of difference schemes for parabolic, hyperbolic and telegraph equations are presented.

2 Stochastic Parabolic Equation

In the papers [6]-[8], the multipoint nonlocal boundary value problem

$$\begin{cases} dv(t) = -Av(t)dt + f(t)dw_t, 0 < t < T, w_t = \sqrt{t}\xi, \xi \in N(0,1), 0 \le t \le T, \\ v(0) = \sum\limits_{j=1}^{J} \alpha_j v(\lambda_j) + \varphi(w_{\lambda_1,\cdots,w_{\lambda_J}}), \sum\limits_{j=1}^{J} |\alpha_j| \le 1, 0 < \lambda_1 < \cdots < \lambda_J \le T \end{cases} \tag{1}$$

for a differential equation in a Hilbert space H with a self adjoint positive definite operator A was investigated. Here and in future:

i. w_t is a standard Wiener process given on the probability space (Ω, F, P).

ii. $f(t)$ is an element of space $M_w^2([0,T], H_1)$ that consists of H_1- value process for which the condition $E\int_0^T \|f(t)\|_{H_1}^2 dt < \infty, H_1 \subset H$ is satisfied.

A single-step difference schemes for the numerical solution of the nonlocal-boundary value problem (1) for a stochastic parabolic equation were presented. The convergence estimates for the solution of the difference schemes were established. In application, the convergence estimates for the solution of difference schemes were obtained for nonlocal-boundary value problems. The theoretical statements for the solution of this difference scheme are supported by numerical examples.

Presently, in the paper [6], 1/2-th order of accuracy Rothe difference scheme for the numerical solution of the Cauchy problem

$$\begin{cases} dv(t) = -A(t)v(t)dt + f(t)dw_t, w_t = \sqrt{t}\xi, \xi \in N(0,1), 0 < t < T, \\ v(0) = 0 \end{cases} \tag{2}$$

for differential equations in a Hilbert space H with self adjoint positive definite dependent in t operators $A(t)$ was constructed. Theorem on convergence estimates for the solution of this difference scheme was established. In applications this abstract result permit us to obtain the convergence estimates for the solution of difference schemes for the numerical solution of initial boundary value problems for parabolic equations.

3 Stochastic Hyperbolic Equation

In the papers **[10],[11], [12]**, the Cauchy problem was investigated

$$\begin{cases} d\dot{v}(t) + Av(t)dt = f(t)dw_t, w_t = \sqrt{t}\xi, \xi \in N(0,1), 0 < t < T, \\ v(0) = \varphi, \dot{v}(0) = \psi \end{cases} \tag{3}$$

for the second order stochastic differential equation in a Hilbert space H with a self adjoint positive definite operator A with $A \ge \delta I$, where $\delta > \delta_0 > 0$.In addition to i) and ii),

iii. For any $z \in [0,T]$, $f(z)$ is an element of the space $M_w^2([0,T],H_1)$, where H_1 is a subspace of H.

Here, $M_w^2([0,T],H)$ denote the space of $H-$valued measurable processes which satisfy

(a) $\phi(t)$ is F_t measurable, a.e. in t and $E\int_0^T \|\phi(t)\|_H dt < \infty$.

iv. φ and ψ are elements of the space $M_w^2([0,T],H_2)$ of H_2-valued measurable processes, where H_2 is a subspace of H.

A two-step difference schemes for solving the Cauchy problem(3) for a stochastic hyperbolic equation were presented. The convergence estimates for the solution of difference schemes for a stochastic hyperbolic equation were established. In applications, the theorems on convergence estimates for the solution of difference schemes for the approximate solution of initial-boundary value problems for stochastic hyperbolic equations were proved. The solutions of the difference schemes were supported by the results of their numerical experiments. Thus, results show that the error is stable and decreases in an exponential manner.

Finally, in paper [13] the two-step difference scheme for the telegraph equation was presented. The convergence estimate for the solution of the difference scheme was established. In applications, the convergence estimates for the solution of difference scheme for the numerical solution of two problems for telegraph equations were obtained. The theoretical statements for the solution of this difference scheme were supported by the results of the numerical experiment.

Finally, I will formulate some problems and future plans.

References

[1] A. Jentzen and P. E. Kloeden, *The numerical approximation of stochastic partial differential equations*, Milan J. of Mathematics **77(1)** (2009) 205–244.

[2] A. Jentzen, *Higher order pathwise numerical approximations of of SPDEs with additive noise*, SIAM Journal on Numerical Analysis **49(2)** (2011) 642–667.

[3] E. Pardoux, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics**3(2)** (1979) 127-167.

[4] S. Peszat and J. Zabczyk,*Nonlinear stochastic wave and heat equations*, Probability Theory and Related Fields**116(3)**(2000) 421-443.

[5] A. Ashyralyev and P. E. Sobolevskii, *New Difference Schemes for Partial Differential Equations*, Operator Theory Advances and Applications, Birkhäuser Verlag, Basel, Boston, Berlin, 2004.

 ISBN: 978-84-608-6082-2

[6] T. Shardlow, *Numerical methods for stochastic parabolic PDEs*, Numerical Functional Analysis and Optimization**20(1-2)** (199) 121–145.

[7] A. Ashyralyev and M.E. San,*An approximation of semigroups method for stochastic parabolic equations*, Abstract and Applied Analysis**Article ID 684248** 24 pages, 2012.

[8] A. Ashyralyev, *On modified Crank-Nicholson difference schemes for stochastic parabolic equation*, Numerical Functional Analysis and Optimization **29(3-4)** 268–282, 2008.

[9] A. Ashyralyev and U. Okur, *Numerical solution of the stochastic parabolic equation with the dependent operator coefficient*, in: Book Series: AIP Conference Proceedings, vol. 1676 24–29, 2015.

[10] A. Ashyralyev, and M. Akat, *An approximation of stochastic hyperbolic equations*, in: Book Series: AIP Conference Proceedings, vol. 1389, 2011.

[11] A. Ashyralyev, and M. Akat,*An approximation of stochastic hyperbolic equations: case with Wiener process*, Mathematical Methods in the Applied Sciences, **36(9)** 1095–1106, 2013.

[12] N. Aggez and M. Ashyralyyewa,*Numerical solution of stochastic hyperbolic equations*, Abstract and Applied Analysis **2012, Article ID 824819, doi:10.1155/2012/824819** 2012.

[13] A. Ashyralyev, and M. Akat, *An approximation of stochastic telegraph equations*, in: Book Series: AIP Conference Proceedings vol. 1479 598–601, 2012.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

BOUNDED SOLUTIONS OF NONLINEAR PARABOLIC EQUATIONS WITH TIME DELAY

Allaberen Ashyralyev[1], Deniz Ağırseven[2] and Burcu Ceylan[3]

[1] *Department of Elementary Mathematics Education, Fatih University, Istanbul, Turkey*

[2] *Department of Mathematics, Trakya University, Edirne, Turkey*

[3] *Department of Computational Science, Trakya University, Edirne, Turkey*

emails: aashyr@fatih.edu.tr, denizagirseven@trakya.edu.tr, brc.cyln@hotmail.com

Abstract

We consider the initial value problem

$$\begin{cases} \frac{du}{dt} + Au(t) = f(u(t), u(t-w)), t \geq 0, \\ u(t) = \varphi(t), -w \leq t \leq 0 \end{cases}$$

in a Banach space E with strongly positive operator A. Theorem on the existence of bounded unique solution of this problem is established. The first and second order of accuracy difference schemes for the solution of the nonlinear parabolic equation with time delay are presented. Theoretical results are supported by numerical experiments.

Key words: delay parabolic equations, difference schemes, initial value problems, bounded unique solution
MSC 2000: 35K60, 65M06

1 Introduction

It is known that, in delay differential equations, the presence of the delay term causes the difficulties in analysis of differential equations, there are a few works that analytic solutions are given. For this reason, the researches in numerical methods recompense the lack of theoretical researches. Especially, one of the main methods used on this area is finite difference method.

Lu [1], studies monotone iterative schemes for finite-difference solutions of reaction-diffusion systems with time delays and give modified iterative schemes by combing the method of upper-lower solutions and the Jacobi method or the Gauss-Seidel method.
Gu and Wang [2], construct a linearized Crank-Nicolson difference scheme to solve a type of variable coefficient delay partial differential equations.
Ashyralyev and Sobolevskii [3], consider the initial-value problem for linear delay partial differential equations of the parabolic type and give a sufficient condition for the stability of the solution of this initial-value problem. They obtain the stability estimates in Hölder norms for the solutions of the problem.
Ashyralyev and Ağırseven [4]-[10], investigated several types of initial and boundary value problems for linear delay parabolic equations. They give theorems on stability and convergence.
In this work, we study existence of bounded unique solution of nonlinear parabolic equation with time delay.

2 Existence Theorem for Bounded Unique Solutions

We consider the initial value problem

$$\begin{cases} \frac{du}{dt} + Au(t) = f(u(t), u(t-w)), t \geq 0, \\ u(t) = \varphi(t), -w \leq t \leq 0 \end{cases} \tag{1}$$

in a Banach space E with strongly positive operator A . The recursive solution of this problem is
$\mathrm{u}_n(t) = e^{-At}u(0) + \int_0^t e^{-A(t-s)}f(u_{n-1}(s), u_n(s-w))ds, n = 1, 2, ...$
$\mathrm{u}_0(t) = \varphi(t)$
The necessary conditions for the existence of unique bounded solution is given by the following theorem.

Theorem 1.
(i) $\varphi(t) \in [0, \infty), \varphi(t) \in D(A)$ and $\|\varphi(t)\|_E \leq M, -w \leq t \leq 0$
(ii)f: E$\times E \longrightarrow E$ be continuous and bounded function, that is $\|\varphi(t)\| \leq \bar{M}$ is satisfied and let L be positive constant, and Lipschitz condition holds according to w.
$\|f(u, w) - f(v, w)\|_E \leq L\|u - v\|_E$
Then Problem (1) has a unique, bounded solution in D.

 ISBN: 978-84-608-6082-2

3 Numerical Results

We consider

$$\begin{cases} \frac{\partial u(t,x)}{\partial t} + \frac{\partial^2 u(t,x)}{\partial x^2} = u(t,x)[u(t-1,x)cosx - u_x(t-1,x)sinx], t \geq 0, 0 < x < \pi \\ u(t,x) = e^{-t}sinx, -1 \leq t \leq 0 \\ u(t,0) = u(t,\pi) = 0, t \geq 0 \end{cases} \tag{2}$$

Using MATLAB programming the numerical results are obtained and error analysis is given by table.

Table 1. Comparison of the errors of different difference schemes in $t \in [n, n+1], n = 0, 1, 2, ...$

Method	N=M=30	N=M=60	N=M=120
first or. of acc. d.s.	0.0064	0.0031	0.0015
second or. of acc. d.s.	4.5864×10^{-4}	1.1212×10^{-4}	2.7577×10^{-5}

References

[1] X. Lu, *Combined iterative methods for numerical solutions of parabolic problems with time delays*, Appl. Math. Comput. **89** (1998) 213–224.

[2] W. Gu, P. Wang, , *A Crank-Nicolson Difference Scheme for Solving a Type of Variable Coefficient Delay Partial Differential Equations*, Journal of Applied Mathematics **Article ID 560567** (2014) 6 pages.

[3] A. Ashyralyev, P. E. Sobolevskii, *On the stability of the linear delay differential and difference equations*, J. Diff. Geom. **6(5)** (2001) 267–297.

[4] A. Ashyralyev, D. Agirseven, *Stability of Parabolic Equations with Unbounded Operators Acting on Delay Terms*, lectronic Journal of Differential Equations **160** (2014) 1–13.

[5] A. Ashyralyev, D. Agirseven, *On source identification problem for a delay parabolic equation*, Nonlinear Analysis: Modelling and Control **19(3)** (2014) 335–349.

[6] A. Ashyralyev, D. Agirseven, *Stability of Delay Parabolic Difference Equations*, Filomat **28:5** (2014) 995–1006.

[7] A.Ashyralyev, D.Agirseven, *Well-posedness of delay parabolic equations with unbounded operators acting on delay terms*, Boundary Value Problems **2014:126** (2014).

[8] A.Ashyralyev, D.Agirseven, *Well-posedness of delay parabolic difference equations*, Advances in Difference Equations **2014:18** (2014).

[9] A. Ashyralyev, D. Agirseven, *On Convergence of Difference Schemes for Delay Parabolic Equations*, Computers and Mathematics with Applications **66(7)** (2013) 1232–1244.

[10] D. Agirsevenn, *Approximate Solutions of Delay Parabolic Equations with the Dirichlet Condition*, Abstract And Appled Analysis **Article Number: 682752** (2012).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Role of Cell Competition in Acquired Chemotherapy Resistance

Piotr Bajger[1], Mariusz Bodzioch[2,3] and Urszula Foryś[3]

[1] *Inter-Faculty Interdisciplinary Doctoral Studies in Natural Sciences and Mathematics, University of Warsaw, Poland*

[2] *Faculty of Mathematics and Computer Science, University of Warmia and Mazury in Olsztyn, Poland*

[3] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*

emails: p.bajger@uw.edu.pl, mariusz.bodzioch@matman.uwm.edu.pl, urszula@mimuw.edu.pl

Abstract

A modification to the tumour growth model proposed by Hahnfeldt et al. in 1999 is considered in this study. The cell population is divided into two subpopulations with different chemotherapy resistances. The influence of competition between the two types of cells on the acquired drug resistance phenomenon is considered. The steady states of the resulting systems are examined. The continuous, indefinite chemotherapy dose is treated as a bifurcation parameter. It is confirmed that higher drug doses may not necessarily extend survival time. The results support the hypothesis that cell competition plays an important role in acquired chemotherapy resistance.

Key words: tumour, chemotherapy, cell competition, mathematical modelling

1 Introduction

One of the major obstacles in a design of effective chemotherapy protocols is acquired drug resistance (ADR). The ADR is a process in which tumour cells become more resistant to the cytotoxic agents over the course of treatment. This phenomenon, attributed to a high mutation rate coupled with large cell division rate of cancer cells, has been the subject of many mathematical models [1].

According to the experimental results regarding the intrinsic drug resistance, a small drug-resistant cell subpopulation may be present in the tumour at the onset of the treatment. Hence the question of optimal chemotherapy dosage naturally arises when ADR is considered. Hostile environment imposed by a high concentration of the cytotoxic agent in blood may promote an outburst of a chemotherapy-resistant tumour cells. Contrary to the maximal dose principle, it is therefore possible that increasing the dosage worsens the outcome of chemotherapy. On the other hand, if the dosage is too small, the treatment may not suffice to effectively inhibit the tumour growth.

In this study, a modification to the framework established by Hahnfeldt et al. in 1999 [2] is proposed in order to take into account the heterogeneity of the malignant cells population. The malignant cells are divided into two subcategories (sensitive and resistant) basing on their resistance to chemotherapy. The ADR is then explained solely by the competition between the two kinds of cancer cells. Potential failures of chemotherapy resulting from unnecessarily high drug dose are discussed.

The major difference between this model and other similar ones proposed in the literature (such as [3]) is the focus on cell competition, rather than cell mutations. The use of Hahnfeldt et al.'s framework also allows for tracking the number of endothelial cells (or, equivalently, the carrying capacity). This is important as the cytotoxic drug is not selective and affects the tumour vasculature as well. Furthermore, it makes the proposed framework viable for including the effects of combined therapy.

2 The model

In 1999, Hahnfeldt et al. [2] proposed a model for tumour growth under angiogenic simulation/inhibition. The model was used to quantify the effects of anti-angiogenic treatment (modelled as a direct death of endothelial cells). Hahnfeldt et al. assumed that the tumour population V follows the Gompertzian-type growth, while the carrying capacity K is the size of the vasculature (related to the number of endothelial cells). This led to the following system of equations:

$$\begin{aligned} \dot{V} &= -\lambda V \ln\left(\frac{V}{K}\right), \\ \dot{K} &= -\mu K + bV - dV^{2/3}K, \end{aligned} \tag{1}$$

where λ is the Gompertzian growth rate of malignant cells, μ is the natural death rate of the endothelial cells, b is the rate at which the vasculature growth is stimulated by the cancer cells and d is a measure of how strongly the cancer cells inhibit the vasculature. See the original article by Hahnfeldt et al. [2] for details on the derivation of these equations.

In order to take into account the heterogeneity of the tumour cells with respect to their

resistance to chemotherapy, the following modification of (1) is proposed in this study:

$$\begin{aligned}
\dot{V}_1 &= -\lambda_1 V_1 \ln\left(\frac{V_1 + \alpha_{12} V_2}{K}\right) - \beta_1 V_1 g(t), \\
\dot{V}_2 &= -\lambda_2 V_2 \ln\left(\frac{V_2 + \alpha_{21} V_1}{K}\right) - \beta_2 V_2 g(t), \\
\dot{K} &= -\mu K + (b_1 V_1 + b_2 V_2) - d\,(V_1 + V_2)^{2/3} K - \beta K g(t),
\end{aligned} \tag{2}$$

where α_{12}, α_{21} are the competition coefficients between the two types of cancer cells, β_1, β_2, β measure how different cell types are affected by the chemotherapy and $g(t)$ is the concentration of the cytotoxic drug. Parameters $\lambda_1, \lambda_2, b_1, b_2, \mu$ are analogous to their equivalents from (1).

It is assumed that the treatment starts upon tumour detection when the total volume of the tumour is $300mm^3$. It is further assumed that at the onset of therapy there are many more sensitive cells than resistant ones. The carrying capacity at the detection point is $650mm^3$, the same as used in the original work. Hence the initial conditions are as follows: $V_1(0) = 280, V_2(0) = 20, K(0) = 650$.

The standard MATLAB solver ode45 was used with error tolerance equal to 10^{-5}. The values of parameter used in all of the simulations are listed in Table 1. The parameters whose values are not listed differ between the simulations. These values can be found in figure descriptions.

It is assumed here that the first type of cells (V_1) is more sensitive to the chemotherapy than the second type (V_2), i.e. $\beta_1 > \beta_2$. At the onset of treatment, the V_2 resistant subpopulation is much smaller than the V_1 sensitive subpopulation. This is only possible, if in the absence of therapy the sensitive cells outcompete the resistant ones, i.e. $\alpha_{21} > \alpha_{12}$ [4].

3 Results

In order to be able to obtain any analytic results and examine the system behaviour, in the following section the chemotherapy is approximated by taking $g(t) = g = const$.

Under this assumption, it can be shown that System (2) has at most three steady states, noting that 0 is not an admissible point. The steady states are denoted by $S_i^* = \{V_1^{i*}, V_2^{i*}, K^{i*}\}$ for $i = 1, 2, 3$.

The first steady state S_1^* corresponds to the situation in which the chemotherapy-sensitive cells outcompete the resistant type. It exists provided that $b_1\,\mathrm{e}^{-\frac{\beta_1}{\lambda_1}g} - \beta g > 0$ and its coordinates $S_1^* = \{V_1^{1*}, 0, K^{1*}\}$ are given by:

$$V_1^{1*} = \left(\frac{b_1\,\mathrm{e}^{-\frac{\beta_1}{\lambda_1}g} - \beta g - \mu}{d}\right)^{3/2} \quad \text{and } K^{1*} = V_1^{1*}\,\mathrm{e}^{\frac{\beta_1}{\lambda_1}g}.$$

 ISBN: 978-84-608-6082-2

This steady state is stable if $(\beta_1\lambda_2 - \beta_2\lambda_1)g < \lambda_1\lambda_2 \ln \alpha_{21}$.

Similarly, the steady state S_2^{2*} at which the resistant cells dominate over the sensitive ones exists provided that $b_2\, \mathrm{e}^{-\frac{\beta_2}{\lambda_2}g} - \beta g > 0$ and its coordinates $S_2^* = \{0, V_2^{2*}, K^{2*}\}$ are given by:

$$V_2^{2*} = \left(\frac{b_2\, \mathrm{e}^{-\frac{\beta_2}{\lambda_2}g} - \beta g - \mu}{d} \right)^{3/2} \quad \text{and } K^{2*} = V_2^{2*}\, \mathrm{e}^{\frac{\beta_2}{\lambda_2}g}.$$

It is stable if $(\beta_1\lambda_2 - \beta_2\lambda_1)g > -\lambda_1\lambda_2 \ln \alpha_{12}$.

The last, positive steady state S_3^* at which both cell types coexist exists if

$$\alpha_{12} < \mathrm{e}^{-\frac{\beta_1 g}{\lambda_1} + \frac{\beta_2 g}{\lambda_2}},\ \alpha_{21} < \mathrm{e}^{-\frac{\beta_2 g}{\lambda_2} + \frac{\beta_1 g}{\lambda_1}},$$
$$b_1 \left(\mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} - \alpha_{12}\, \mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} \right) + b_2 \left(\mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} - \alpha_{21}\, \mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} \right) > (\mu + \beta g)(1 - \alpha_{12}\alpha_{21}).$$

The coordinates of S_3^* are given by

$$V_1^{3*} = \frac{\mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} - \alpha_{12}\, \mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}}}{1 - \alpha_{12}\alpha_{21}} K^{3*},$$
$$V_2^{3*} = \frac{\mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} - \alpha_{21}\, \mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}}}{1 - \alpha_{12}\alpha_{21}} K^{3*},$$
$$K^{3*} = \frac{\left[b_1 \left(\mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} - \alpha_{12}\, \mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} \right) + b_2 \left(\mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} - \alpha_{21}\, \mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} \right) - (\mu + \beta g)(1 - \alpha_{12}\alpha_{21}) \right]^{3/2}}{d^{3/2} \left(\mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} - \alpha_{12}\, \mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} + \mathrm{e}^{-\frac{\beta_2 g}{\lambda_2}} - \alpha_{21}\, \mathrm{e}^{-\frac{\beta_1 g}{\lambda_1}} \right)}.$$

The conditions for stability of the steady state S_3^* obtained by the Routh-Hurwitz criterion are very complex, hence not provided here. It is, however, possible to show using this criterion that if both cell types are assumed to stimulate the angiogenesis at the same rate (i.e. $b_1 = b_2$), the steady state S_3^* is always stable.

In what follows, the drug dose g will be treated as a bifurcation parameter. A particularly interesting result emerges if $\beta_1\lambda_1 - \beta_2\lambda_1 > 0$. The qualitative properties of the system diagram are then determined by the values of the competition coefficients α_{12} and α_{21}.

In Figure 1, the tumour size at the steady states is plotted against the bifurcation parameter g for different values of α_{12} and α_{21}. The tumour size is defined as a weighted average $\sigma_1 V_1 + \sigma_2 V_2$, where $\sigma_1 + \sigma_2 = 1$. The rationale behind taking a weighted average rather than simply an overall tumour volume $V_1 + V_2$ is that it is more desirable for the tumour to be composed of the sensitive cells rather than resistant ones. Hence to penalise tumour resistance σ_2 should be taken to be greater than 0.5. In other words, the aim of

Name	λ_1	λ_2	μ	b_1	b_2	d
Unit	1/day	1/day	1/day	1/day	1/day	$\text{day}^{-1}\text{vol}^{-2/3}$
Value	0.192	0.192	0	5.85	5.85	0.00873
Name	β_1	β_2	β			
Unit	$\text{day}^{-1}\text{conc}^{-1}$	$\text{day}^{-1}\text{conc}^{-1}$	$\text{day}^{-1}\text{conc}^{-1}$			
Value	0.15	0.1	0.05			

Table 1: Parameters for the model. The values of parameters $\lambda_1, \lambda_2, \mu, b_1, b_2, d$ are taken from [2], while the values of β_1, β are taken from [5]. The value of β_2 was chosen arbitrarily.

the treatment may not necessarily be to just maximise cell death, but also to minimise the ADR effect.

In Figures 1a,b it is visible how increasing the drug dose leads to a smooth transition from a tumour consisting of chemotherapy-sensitive cells to a one composed mainly of the resistant cells. Figure 1a shows that for $\sigma_1 = \sigma_2 = 0.5$ the steady state tumour size decreases as the drug dose g increases. However, if the resistant cells population is penalised by taking $\sigma_1 = 0.4, \sigma_2 = 0.6$ (Figure 1b), an increase in dose may result in an increase in tumour size.

Figures 1c,d show a hysteresis loop. This is important from a point of view of effective chemotherapy planning and may potentially be dangerous for the patient. It can be seen that increasing the drug dose does not necessarily mean a decrease in tumour size. Furthermore, if the administrated dose exceeds a critical value at which S_1^* loses stability, the tumour transits to the "resistant" steady state. The harmful effects of this transition are even more apparent when the resistant cells population is penalised ($\sigma_2 = 0.6$). Reversing this transition is then only possible if the drug dose is decreased below the value at which S_2^* gains stability.

The danger associated with the hysteresis loop is visualised in Figure 2. Two chemotherapy protocols are considered. In Figure 2a the drug concentration admits a value $g = 1.5$ and does not change in time. In Figure 2b this dose is increased to 2.5 between days 60 and 100 and equal to 1.5 at all other times. It is visible how the tumour transits to the "resistant" state and simple reversal of the drug dose increase was not enough to transit back to the "sensitive" state.

In light of the above, a potential failure of chemotherapy may be therefore associated with an unnecessary increase in dosage. To illustrate this phenomenon, the System (2) was solved numerically. The results of these simulations are discussed below.

Figure 3 shows the survival time plotted against the drug dose. Survival time is defined to be the time until the overall tumour volume $V_1 + V_2$ reaches a critical value V_{crit}. Three values of V_{crit} were tested. All three plots show a slow initial increase in survival time. This is then followed by a spike – a drug dose corresponding to a good survival/cytotoxicity ratio. A decrease in the survival time is then visible until the chemotherapy dose is large

 ISBN: 978-84-608-6082-2

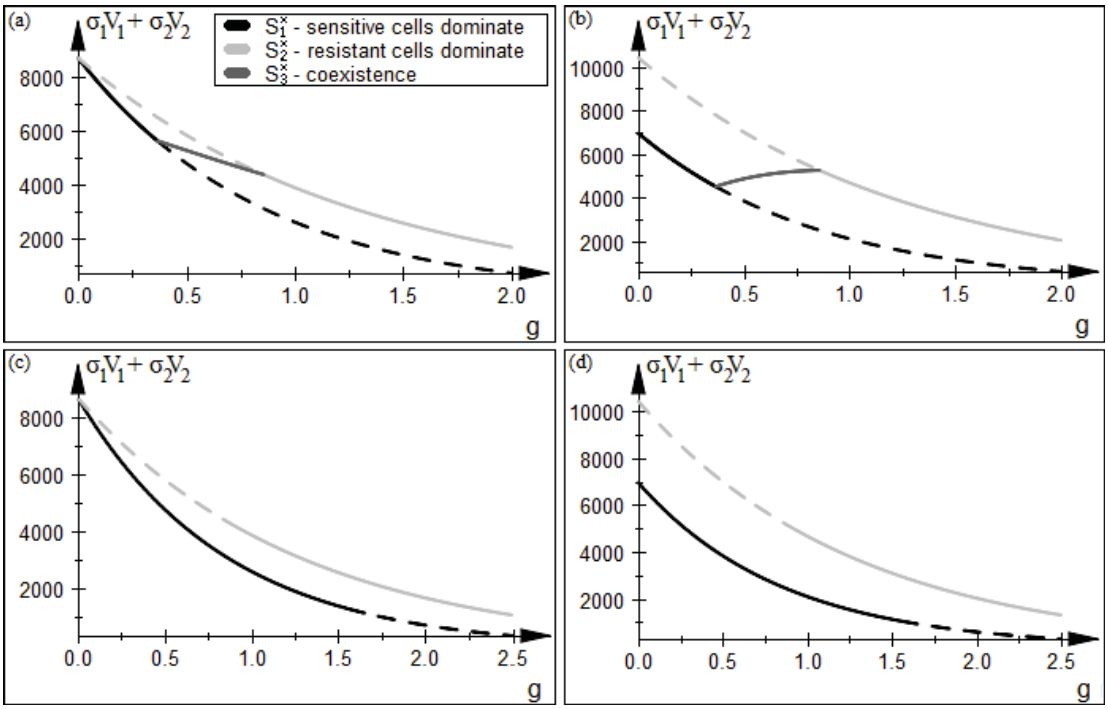


Figure 1: Bifurcation diagram for the system (2). Weighted average of the volumes of the two cell types at the steady states are plotted against a chemotherapy dose g. Solid lines denote stability, while dashed ones instability. The parameter values are: (a) $\alpha_{12} = 0.8, \alpha_{21} = 1.1, \sigma_1 = 0.5, \sigma_2 = 0.5$, (b) $\alpha_{12} = 0.8, \alpha_{21} = 1.1, \sigma_1 = 0.4, \sigma_2 = 0.6$, (c) $\alpha_{12} = 0.8, \alpha_{21} = 1.5, \sigma_1 = 0.5, \sigma_2 = 0.5$, and (d) $\alpha_{12} = 0.8, \alpha_{21} = 1.5, \sigma_1 = 0.4, \sigma_2 = 0.6$.

enough to maintain the tumour volume below the critical value indefinitely.

Figure 4 shows examples of the tumour evolution for different drug doses. As expected, an increase in the dose promotes the growth of the chemotherapy resistant cells. In Figure 4a, the dose chosen is too small to effectively inhibit tumour growth. In Figure 4b, the dose is enough to maintain the tumour size below the critical value for over 200 days until the resistant cells eventually dominate. In Figure 4c, the administrated dose was too large and a quick outburst of resistant subpopulation is present.

4 Discussion

The results obtained for continuous indefinite chemotherapy support the hypothesis that cell competition plays an important role in ADR. Interesting dynamics which arises for

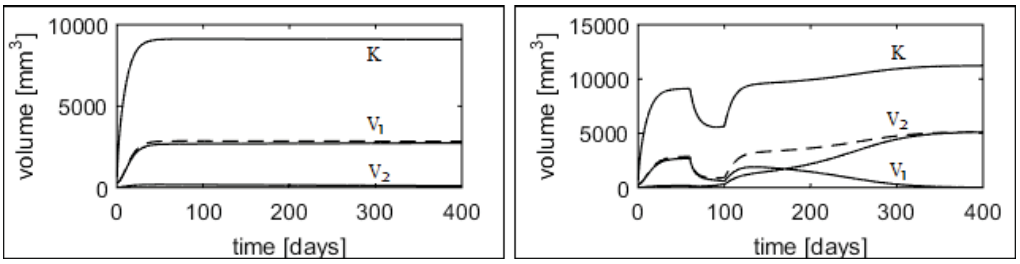


Figure 2: Numerical solutions to System (2) for different chemotherapy protocols. The values of the competition coefficients are $\alpha_{12} = 0.8, \alpha_{21} = 1.5$. The drug doses are (a) $g(t) \equiv 1.5$ and (b) $g(t) = 2.5$ for $t \in [70, 100]$ and $g(t) = 1.5$ at all other times. Dashed line shows the total tumour volume $V_1 + V_2$.

certain values of the competition coefficients suggests that cell competition should not be completely disregarded in modelling of the ADR effect.

The qualitative differences in the dynamics of the system for different values of the competition parameters α_{12}, α_{21} are summarised in Figure 1 by means of a bifurcation diagram.

If the sensitive cells do not exert a large influence over the resistant cells (i.e. α_{21} close to 1), the transition from the "sensitive" tumour to the "resistant" one is smooth and reversible by adjusting the chemotherapy dose. It should be noted, however, that even the largest dose $g = 2$ is not enough to maintain the total tumour volume under the desired critical value V_{crit} (Figure 1a,b). Larger drug dose is necessary and the resistant cells have to be killed anyway.

If the sensitive cells, however, are very good at suppressing the growth of the resistant type (e.g. $\alpha_{21} \approx 1.5$), the dynamics is more interesting. The system exhibits hysteresis, as shown in Figure 2b,c. As the sensitive cells die out in course of the treatment, they can no longer keep the resistant cells at bay. This results in an outburst of the resistant subpopulation and a transition to the "resistant" steady state. Due to the hysteresis, a decrease in the chemotherapy dose may not be enough to return to the "sensitive" steady state. This has potentially dangerous implications for the patient, as larger chemotherapy doses are necessary to eradicate the resistant tumours.

In line with similar results obtained by other authors [3], Figure 3 suggests that the increase in chemotherapy dose does not necessarily imply an increase of the survival time. In fact the doses with a good survival/cytotoxicity ratio will be the ones below the value of g at which the "sensitive" steady state loses stability, i.e. $g < -\lambda_1\lambda_2 \ln(\alpha_{21}) / (\beta_1\lambda_2 - \beta_2\lambda_1)$.

Although the model described in this study supports the hypothesis that cell competition significantly contributes to the ADR effect, the authors are aware of its current limitations. In particular, although a spike in survival time is present for intermediate drug

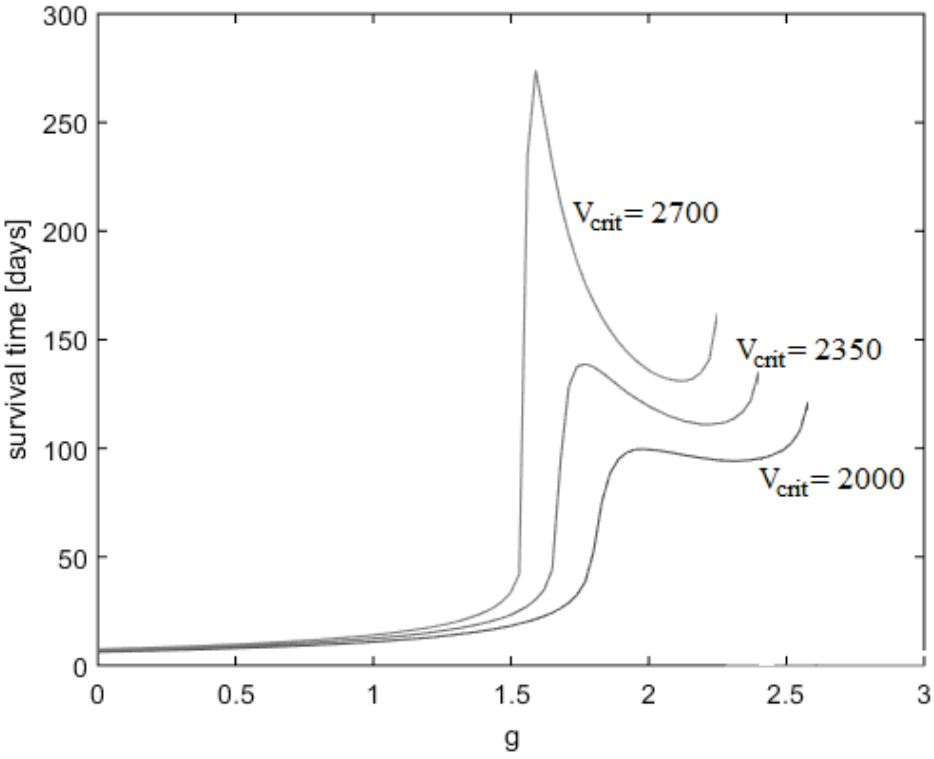


Figure 3: Survival time plotted against a chemotherapy dose for different values of V_{crit}. Each plot ends when the survival time becomes infinite (tumour volume never reaches the critical value).

dose, a dose enough to eradicate the tumour completely is predicted to be not much higher. The exact quantitative results are to be treated with caution, nevertheless an absolute cure with no much increase in cytotoxicity may be considered even better for the patient. It is, however, a commonly accepted theory that rapid cell mutations are also responsible for the ADR. These mutations are likely to increase a drug dose required for a complete cure. It would therefore be beneficial to include such mutations into the model, as suggested in [3] in a similar setting.

What is more, the potential of Hahnfeldt et al. framework is not fully capitalised in the current form of the model. The fact that the size of the tumour vasculature is also being modelled makes it easier to include effects of the anti-angiogenic treatment and hence combined therapy. The subtle interplay between these two kinds of treatment is yet to be fully understood. Combined therapy could be considered using modifications of Hahnfeldt

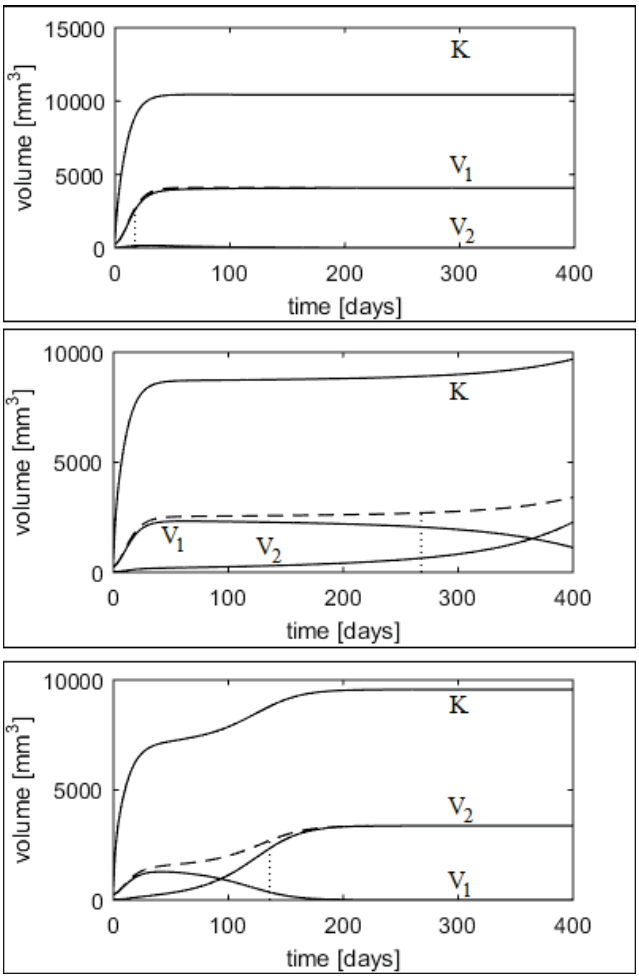


Figure 4: Examples of three solutions to System (2) for different values of g. The competition coefficients are $\alpha_{12} = 0.8, \alpha_{21} = 1.5$ and the drug doses are (a) $g = 1.2$, (b) $g = 1.6$, (c) $g = 2$. Dashed line shows the total tumour volume $V_1 + V_2$. Dotted vertical line denotes the survival time for $V_{crit} = 2700$.

et al. model proposed in [5, 6]. It would be interesting to see what possible implications combined therapy may have on the ADR.

Finally, although a continuous indefinite chemotherapy is particularly suitable for mathematical analysis, it suffers from not being clinically realistic. Investigation of the effects of the drug dose and density by choosing g to be a more realistic pharmacokinetic function is a good next step in the development of the model.

 ISBN: 978-84-608-6082-2

In summary, this work modifies the framework proposed by Hahnfeldt et al. to account for tumour cells with varying chemotherapy resistance. Preliminary results regarding the continuous, constant indefinite chemotherapy protocol support the hypothesis that cell competition plays an important role in the ADR effect. It confirms that spikes in survival time are present for intermediate values of the drug dose. Next steps in model development will include: mutation to resistance, more realistic chemotherapy protocols and inclusion of anti-angiogenic therapy.

Acknowledgements

The second author was partially supported by WCMCS, Warsaw, within PhD Internships programme.

References

[1] J. Foo and F. Michor. Evolution of acquired resistance to anti-cancertherapy. *Journal of Theoretical Biology*, 355:10–20, 2014.

[2] P. Hahnfeldt, D. Panigrahy, J. Folkman, and L. Hlatky. Tumor development under angiogenic signaling: A dynamical theory of tumor growth, treatment response, and postvascular dormancy. *Cancer Research*, 59:4770–4775, 1999.

[3] H. C. Monro and E. A. Gaffney. Modelling chemotherapy resistance in palliation and failed cure. *Journal of Theoretical Biology*, 257:292–302, 2009.

[4] J. D. Murray. *Mathematical Biology I: An Introduction*. Springer, 2007.

[5] J. Poleszczuk and I. Skrzypczak. Tumor angiogenesis model with variable vessels' effectiveness. *Applicationes Mathematicae*, 38:33–49, 2011.

[6] J. Poleszczuk, M. Bondnar, and U Foryś. New approach to modeling of antiangiogenic treatment on the basis of hahnfeldt et al. model. *Mathematical Biosciences and Engineering*, 8:591–603, 2011.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Inverse estimation of terminal connections in the cardiac conduction system

Fernando Barber[1], Miguel Lozano[1], Ignacio García-Fernández[1] and Rafael Sebastian[1]

[1] *Department d'Informàtica, ETSE, Universitat de València*

emails: fernando.barber@uv.es, miguel.lozano@uv.es, ignacio.garcia@uv.es, rafael.sebastian@uv.es

Abstract

Modeling the cardiac conduction system is a challenging problem in the context of computational cardiac electrophysiology. Its main structure in the ventricles is the Purkinje system, which is responsible for triggering electrical activation and subsequent heart contraction. Since it cannot be visualized by means of medical imaging techniques, it is commonly neglected or inversely estimated. In this paper we present an algorithm that is capable to estimate the location of the Purkinje system terminals from a set of invivo clinical measurement where the activation time is known. We evaluate the performance of the algorithm for different scenarios. Results show that the method is capable of locating most of the terminals in simple scenarios. When a scenario is complex, the method can locate the terminals with major impact in the overall activation map. Furthermore, the low values associated to the mean square error indicate that solutions provided are useful to simulate a patient ventricular activation map.

Key words: Cardiac Conduction System, Purkinje Tree, Non-statistical estimation, Approximation

1 Introduction

Computational modelling of the heart aims at helping to understand the complex structure and function of the heart in health and disease. The construction of realistic computational cardiac models that can be personalized to a patient is challenging and requires the fusion of disparate medical and biological data [6, 3]. On the one hand, the function of the heart is modeled from ex-vivo experiments at cell and tissue scales, which are usually represented

by sets of differential equations that describe their coupled nonlinear behaviour. On the other hand, a personalized 3D heart geometry can be obtained from segmentation of medical images acquired from computed tomography (CT) or magnetic resonance imaging (MRI). Subsequently, segmented hearts are meshed in 3D to define a fine computational domain in which finite element methods can be used to simulate the heart function. However, the heart is highly complex and inhomogeneous, and not all the cardiac structures and tissue properties can be personalized from medical images due to imaging resolution limits. Among these structures there is the cardiac conduction system (CCS) or Purkinje tree (in the ventricles), which is responsible for the synchronized activation of the cardiac muscle that triggers the heart contraction. In a healthy human heart, the CCS functions as a 'highway' placed in the inner cardiac surface, where the electrical signals travel fast up to the Purkinje-myocardial junctions (PMJs). At each PMJ the impulse enters the 'working' contractile myocardium, which slowly propagates the electrical signal as a wavefront activating the heart tissue. Once a given cell has been activated, it cannot be activated again during the same heart cycle by a second electrical wavefront. As a result, not all PMJs are *effective*, in the sense that many propagate their signal to active cells and make no practical effect, or their contribution is masked by other surrounding PMJs.

The structure of the Purkinje tree and the location of the PMJs cannot be obtained from any in-vivo imaging technique. This fact has led researchers to the construction of generic population-based computational models of the CCS in the ventricles [2, 8]. The reader is referred to [8] for a review on different techniques commonly used to build computer models of the CCS. Recently, some techniques have been developed to infer the location of PMJs and structure of a Purkinje tree from electro-anatomical maps (EAMs) acquired in-vivo in the clinic [4, 9, 7] during radio frequency ablation (RFA) interventions. EAMs are one the the few invivo clinical source of information that can be directly used to obtain the electrical function of the heart. From EAMs, the local activation times (LAT) can be measured using catheters at different locations on the inner surface of the ventricles.

In the works of Palamara et al. [7] and [9], a mathematical method to define positions of PMJs is presented that reduces simulated activation errors given a precomputed generic Purkinje tree structure. In [4], the location of PMJs is determined from singularity points at highly dense activation maps obtained from simulations. None of the techniques try to obtain the real location of the PMJs, but a distribution of them that is coherent with tissue global activation sequences. The real location of all the electrical triggers in the heart which can include PMJs or pathological tissue, such as ectopic activity, is valuable for both constructing realistic models of the heart and planning RFA interventions.

Our goal is to perform an inverse estimation of the location and activation time of PMJs on the inner surface of the heart, given a set of scattered measurements randomly distributed, similarly to those obtained in EAMs. In this work, we present an estimation method, and analyze it on several synthetic scenarios. The structure of the Purkinje tree, as

observed macroscopically in animal studies [1], forms a complex interconnected network. To represent this network we build tree structures recursively, creating two perpendicular child branches at the end of each branch with decreasing length and locating the PMJs at the leaf nodes of the tree. We build a series of scenarios using different Purkinje trees with an increasing number of branches and PMJs density, and present the estimation capabilities. The approximation of the structure of the Purkinje tree from the estimated PMJs is out of the scope of this study.

2 PMJ estimation method

To show and test the methodology we use a simplified scenario that takes the following assumptions: the cardiac tissue is represented by a 2D, euclidean domain, $\Omega \subset \mathbb{R}^2$; the signal propagation is considered isotropic and; the propagation velocity constant. According to the previous description of a Purkinje tree, we assume that the signal enters cardiac tissue through a set $\mathcal{S}$ of n PMJs, with locations $\mathbf{s}_k \in \Omega, k = 1, \ldots, n$. The activation time of node k will be denoted as $\tau_k \in \mathbb{R}$.

Given a point $\mathbf{x} \in \Omega$, its local activation time (from now on LAT) will be the earliest arrival time of the signal from the source nodes, i.e. PMJs, and is given by

$$t(\mathbf{x}) = \min_k \left(\tau_k + \frac{\|\mathbf{x} - \mathbf{s}_k\|}{v} \right), \tag{1}$$

where v is the propagation velocity of the signal through the cardiac tissue. Figure 1 (a) shows a simple scenario with two PMJs represented with solid circles, and several points with measurements (LATs) represented with crosses. Three of the measurement points displayed with LATs t_{i1}, t_{i2} and t_{i3}, were activated by the same PMJ (indicated with arrows).

Since the Purkinje tree cannot be observed, PMJs will be considered as unknown, both in their number and location. Our goal is to estimate the location of the set of *effective* PMJs that produces the observed measurements, which are the activation times at m given locations. Thus, we state our problem as

Problem 2.1 (PMJ estimation) *Given a set $\mathcal{P}$ of tuples $(\mathbf{p}_l, t_l) \in \Omega \times \mathbb{R}, l = 1, \ldots, m$, where t_l is the known activation time at $\mathbf{p}_l$, find a set $\mathcal{F}$ of estimated PMJs and associated activation times $(\hat{\mathbf{s}}_i, \hat{\tau}_i),\ i = 1, \ldots, r$, that minimises the error function*

$$E = \frac{1}{m} \sum_{l=1}^{m} (t_l - \hat{t}_l)^2 \tag{2}$$

where $\hat{t}_l$ is the estimated activation time defined by (1), for $\mathbf{x} = \mathbf{p}_l$ and the min *function ranging in $(\hat{\mathbf{s}}_i, \hat{\tau}_i),\ i = 1, \ldots, r$.*

 ISBN: 978-84-608-6082-2

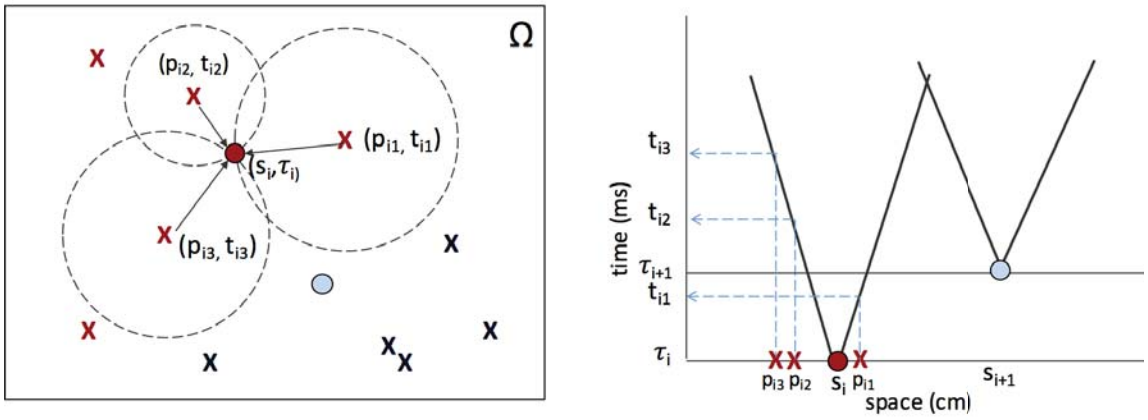


Figure 1: PMJs and measurement points. Spatial representation in a 2D plane with spatial location of PMJs and measurement points (left) and temporal representation using cones (right).

Let us consider a PMJ with spatial coordinates $\mathbf{s} \in \Omega$ and activation time τ. And let us consider a point $\mathbf{p}$ that is activated by the propagation of the signal from $\mathbf{s}$. The activation time t of point $\mathbf{p}$ must meet the equation

$$\|\mathbf{p} - \mathbf{s}\| = v(t - \tau). \tag{3}$$

Equation (3) defines the positive half of a cone (see Figure 1 right) with its vertex in $(\mathbf{s}, \tau)$; thus, for any point activated by $\mathbf{s}$ the point $(\mathbf{p}, t) \in \mathbb{R}^3$ belongs to that cone. Given three points activated by the same PMJ, if a cone can be found in $\mathbb{R}^3$ containing their coordinates in the form $(\mathbf{p}_l, t_l)$, then PMJ coordinates $(\mathbf{s}, \tau)$ must be located at the cone's vertex.

To construct a set of candidate PMJs, we are going to use this property. As a first step in our method we build a Delaunay triangulation for the set of measured points, considering their spatial coordinates. Such a triangulation is the construction of an irregular mesh that takes the points as vertexes and in which all the faces are triangles. Figure 2 shows the Delaunay triangulations for the smallest and largest sizes of $\mathcal{P}$, considered in our experiments.

Let's call

$$\mathcal{T} = \{\{k_1, k_2, k_3\} : \triangle \mathbf{p}_{k_1}\mathbf{p}_{k_2}\mathbf{p}_{k_3} \text{ belongs to the Delaunay triangulation.}\},$$

given a triangle $i = \{i_1, i_2, i_3\} \in \mathcal{T}$, formed by the measurement points, $\mathbf{p}_{i_1}$, $\mathbf{p}_{i_2}$ and $\mathbf{p}_{i_3}$, with activation times, t_{i_1}, t_{i_2} and t_{i_3}, we look for a solution $\mathbf{f}_i = (\hat{\mathbf{s}}_i, \hat{\tau}_i) \in \mathbb{R}^3$ of the system

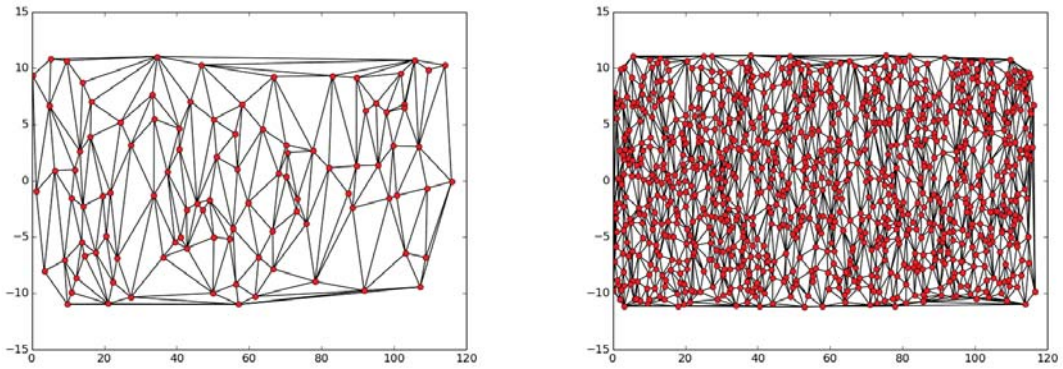


Figure 2: Two different measurement point sets, $\mathcal{P}$, and their corresponding Delaunay triangulation. (Left: $|\mathcal{P}| = 100$, right: $|\mathcal{P}| = 1000$).

of nonlinear equations

$$\begin{aligned} \|\mathbf{p}_{i_1} - \hat{\mathbf{s}}_i\| - v(t_{i_1} - \hat{\tau}_i) &= 0 \\ \|\mathbf{p}_{i_2} - \hat{\mathbf{s}}_i\| - v(t_{i_2} - \hat{\tau}_i) &= 0 \\ \|\mathbf{p}_{i_3} - \hat{\mathbf{s}}_i\| - v(t_{i_3} - \hat{\tau}_i) &= 0 \end{aligned} \tag{4}$$

Moreover, a valid PMJ will have an activation time earlier than the LAT of the three measurements points that generate it, namely

$$\tau_i < t_{i_j}, \quad j = 1, 2, 3. \tag{5}$$

To extract a set $\mathcal{F}$ containing valid PMJs belonging to the hidden tree, we have designed Algorithm 1. For each triangle i included in the Delaunay triangulation $\mathcal{T}$ we look for a valid tentative PMJ, $\mathbf{f}_i$. This step, corresponding to line 5 in the algorithm, consists of the solution of the system (4). This system is solved by means of a standard solver for systems of nonlinear equations. In particular, we use a modification of the Powell hybrid method as implemented in MINPACK [5].

Since the solver can return false PMJs, $\mathbf{f}_i$ needs to be validated. The validation function, called in line 6 of the algorithm, checks three conditions and only accepts the candidate PMJ if all of them are met. The first condition checked is the verification that the candidate PMJ is consistent with the three measurement points that have generated it. This test is done by checking if the solver has converged to a solution of (4) and verifying the conditions (5).

The second condition checked is whether the estimated PMJ is inside the Delaunay

Algorithm 1 PMJs estimation

```
Input → 𝒫 {Set of measurement points with activation time}
ℱ ← ∅ {PMJs}
𝒯 ← delaunay(𝒫) {Build a Delaunay triangle mesh}
for all tri ∈ 𝒯 do
    f_i ← find_source(tri) {The solver finds out a local solution from the triangle tri }
    if is_valid(f_i) then
        ℱ ← ℱ ∪ f_i
    end if
end for
Output ← ℱ {Set of estimated PMJs }
```

triangulation $\mathcal{T}$, otherwise it is considered to be outside the domain of the problem and is disregarded.

The third condition requires that the estimated PMJ is compatible with the backward eikonal problem associated to the triangle where the estimated PMJ is located in. This criterion has already been used by other authors [7] and states that a measurement point cannot activate later than the traveling wavefront produced by the closest PMJ. Thus, we request that the vertexes of triangle k containing $\mathbf{f}_i = (\hat{\mathbf{s}}_i, \hat{\tau}_i)$ meet the condition

$$t_{k_j} \leq \frac{\|\hat{\mathbf{s}}_i - \mathbf{p}_{k_j}\|}{v} + \hat{\tau}_i + \epsilon, \quad j = 1, \ldots, 3 \tag{6}$$

where ϵ is a tolerance parameter that accounts for possible numerical errors.

The estimation problem, as stated in Problem 2.1, admits a trivial solution consisting on placing a PMJ at every measurement point. Obviously, this is an undesired solution. However, by construction the PMJs will be placed at locations different than those of the measurement points, in most if not all cases.

3 Performance Evaluation

To evaluate our approach, we have generated several scenarios consisting of a simulated Purkinje tree structure along with a set of measurement points uniformly distributed. In this section we describe the experimental methodology used and discuss the results.

3.1 Methodology

The test trees that represent the cardiac conduction system are built procedurally. We use a recursive algorithm that, at every level, creates two new branches at the end of each branch that was built in the previous level. Every new branch is perpendicular to the parent

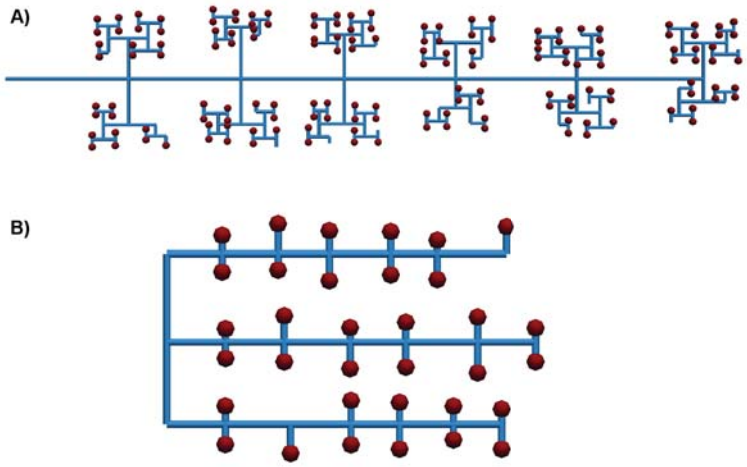


Figure 3: Example of generated trees. A) 1 main branch with depth 2 (B1D6), B) 3 main branches with depth 2 (B3D2)

branch. The leaf nodes of the tree are the PMJs, which are sources of electrical signal in the myocardium. The density of PMJs is indirectly controlled by the depth of the branch recursion. Branch lengths are generated following a normal distribution with parameters obtained from [8]. Figure 3 shows two examples of trees used in our experiments.

We generate two types of scenarios: i) trees with a single main branch, and ii) trees with three main branches connected as depicted in Figure 3. For each type, we consider 3 different recursion depths for generating the tree subbranches: depth 2, depth 4 and depth 6. We will label every configuration with letter B followed by the number of branches plus letter D followed by the depth used to generate it. Figure 3 (a) shows a tree with one branch and recursion depth of 6 (B1D6). Figure 3 (b) shows a tree with three branches and a recursion depth of 2 (B3D2).

Given a simulated Purkinje tree, we firstly set a number of measurement points in the domain varying from 100 to 1000 in steps of 100, distributed uniformly. Secondly, the LAT at each measurement point is computed, by propagating the electrical signal along the tree, up to the PMJs and from the PMJs to each sensor through the shortest path, using Equation (1). The signal within the tree propagates around three times faster than on the local domain. The experimental process can be summarized as follows: i) Generate an artificial tree with a set of PMJs as source points, $\mathcal{S}$; ii) Generate $\mathcal{P}$ measurement points distributed uniformly; iii) Propagate the signal from $\mathcal{S}$ to $\mathcal{P}$ and set the corresponding activation times

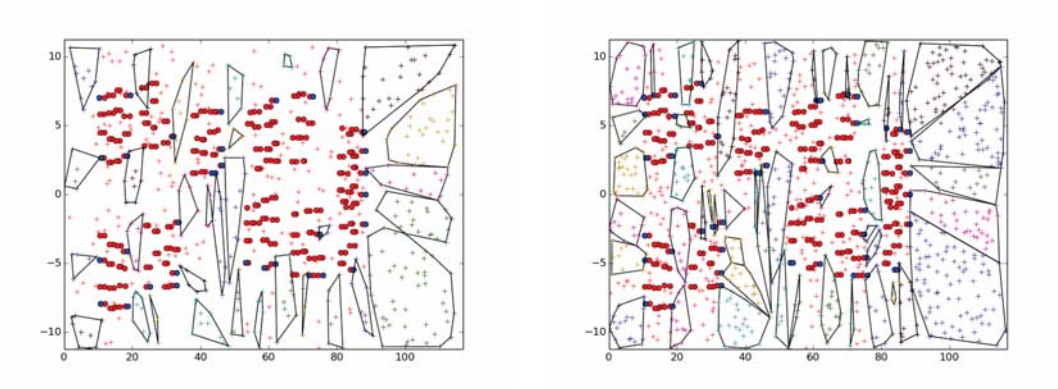


Figure 4: Solutions provided for scenario B3D6 with 500 measurement points (a) and 1000 measurement points (b).

for every point in $\mathcal{P}$; iv) Run the algorithm provided and obtain an estimation $\mathcal{F}$; Compare the solution $\mathcal{S}$ with the estimation $\mathcal{F}$.

In Figure 4 the results of the algorithm can be observed for an scenario, with two different measurement points densities. In the figure, measurement points are represented by crosses. The blue dots represent the PMJs that have been found, while the red dots represent PMJs that have not been found in the solution by our algorithm. The set of all points which are activated by the same PMJs is represented by a polygon containing them.

To compare the generated solution $\mathcal{S}$ (considered as the true unknown cardiac conduction system) and the estimation $\mathcal{F}$ provided by our algorithm, we consider two error measurements. First, we consider the number of PMJs found, and compare it to the actual number of PMJs in the scenario. Second, we measure the estimated LATs, generated by the propagation of the signal from the PMJs in $\mathcal{F}$, and compute the mean square error (Eq. (2) in Problem 2.1). Since the trees and the measurement points are generated randomly, we build a sample of 100 different instances for every scenario and compute the mean and standard deviation of these performance indicators.

3.2 Results

The results of the simulation study, including all the scenarios, are summarized in Figure 5. The number of sources, i.e. PMJs, estimated for the scenario with a single main branch is displayed in Figure 5 (a), while Figure 5 (b) shows the PMJs estimated using three main branches. Every row represents a different recursion depth, for depths 2, 4 and 6. Each

plot shows the number of PMJs in the scenario (F_REAL), the number of PMJs that can be truly estimated (F_DET) or effective sources, the total number of PMJs estimated by the algorithm (F_EST) and finally, the number of PMJs correctly estimated (F_OK).

In Figure 5, the difference between the values of the estimated (F_EST) and actual (F_OK) number of PMJs indicates that some false positives are obtained (estimated PMJs in $\mathcal{F}$ that were not in $\mathcal{S}$). This becomes specially relevant when increasing the number of measurement points. They can appear, for instance, due to a low solver precision, since sometimes the solver provides a set of very close PMJs that should have been collapsed into a single one. Another situation where we have found false positives comes from points that are not activated by the same real PMJ, but generate a feasible candidate. Although several validation tests are performed during the algorithm, as discussed in Section 2, the results show that the validation function is not enough to avoid all of them.

In scenarios B1D2, B1D4 and B3D2 the algorithm obtains nearly all PMJs when it has enough measurement points. However, when the density of PMJs increases the problem becomes more complex, since these points tend to be clustered, as Figure 4 shows.

Figure 6 (a) shows the number of PMJs estimated when increasing the number of actual PMJs in the scenario, while maintaining a fixed number of 1000 measurement points. In the figure, we have combined all the scenarios in increasing order of PMJs so each point in the plot corresponds to a specific scenario (B1D2, B3D2, B1D4, B3D4, B1D6 and B3D6). In this figure we can also notice the aforementioned problems when increasing the density of PMJs. In the first four scenarios, the number of estimated PMJs increases with the total number of PMJs. However this does not occur in the last two scenarios (B1D6, B3D6) where the number of found PMJs decreases. This behaviour is also amplified because we have set a fixed number of measurement points. As a consequence, the ratio between measurement points and PMJs decreases, making it more difficult to build a complete estimate.

Despite these apparent limitations, Figure 6 (b) shows how the mean squared error (MSE) associated to all the tested scenarios decreases rapidly with the measurement points. This behavior of the MSE reveals that the set of correctly estimated PMJs are the most significant ones. The solution shown in Figure 4 gives a hint on the reason for this; the PMJs which are properly detected are close to the border of the region occupied by the tree, while the inner PMJs are those missed by the algorithm. The signal emitted by these inner points is quickly masked by the points at the border of the tree and they are mainly non-effective PMJs. From these results, we conclude that our algorithm is capable of finding most of the effective nodes of the Purkinje system.

4 Conclusions and future work

We have presented a methodology to estimate, from electro-anatomical map (EAMs) samples, the location and activation time of the sources of electrical activation, known as PMJs,

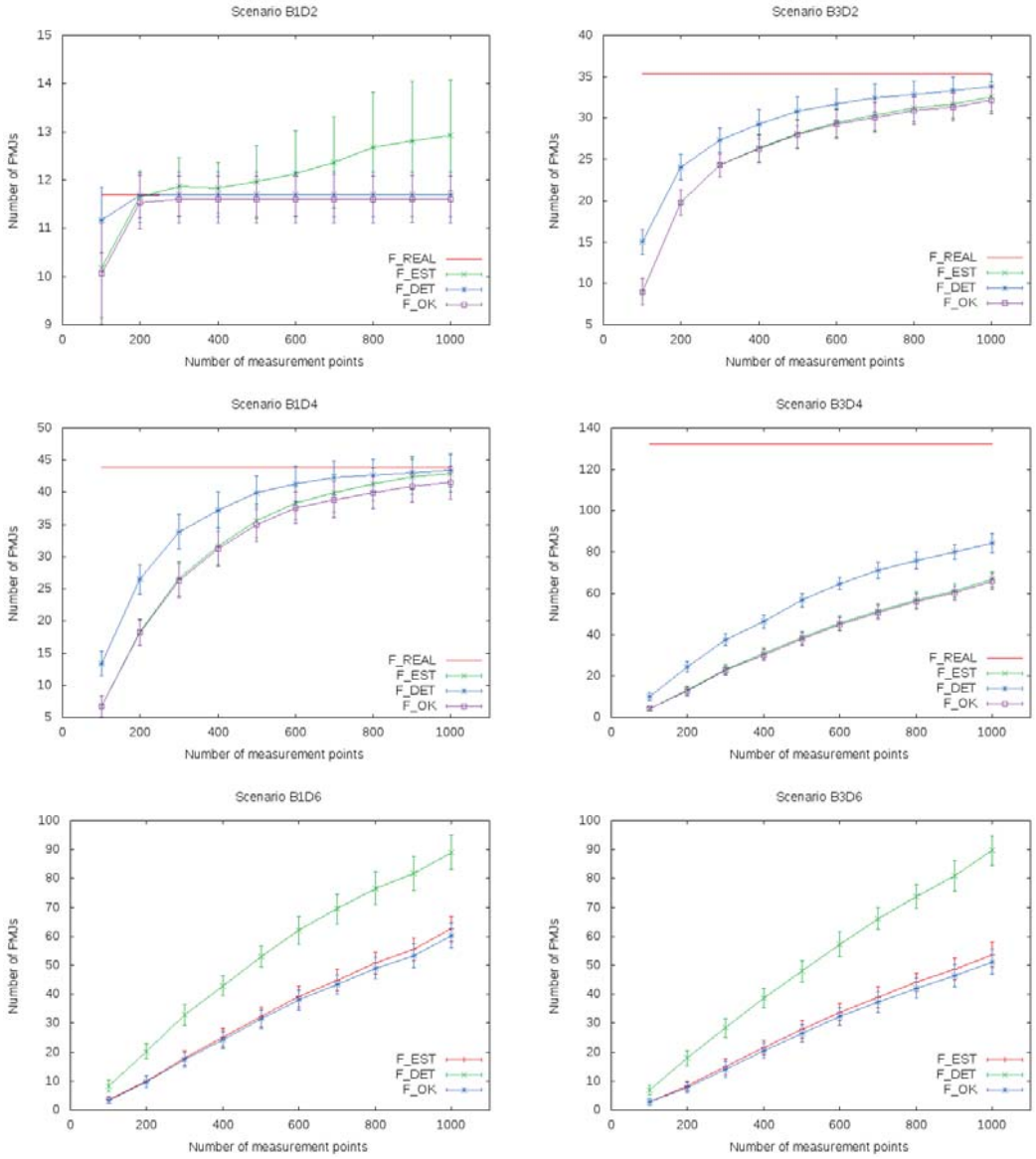


Figure 5: Estimation results for a single main branch (a), and for three main branches (b). Every row corresponds to a different recursion depth. The used recursion depths are (from top to bottom) 2, 4 and 6. Figures are not at the same scale, since the actual number of PMJs varies for the different scenarios. See the text for details on the figure contents.

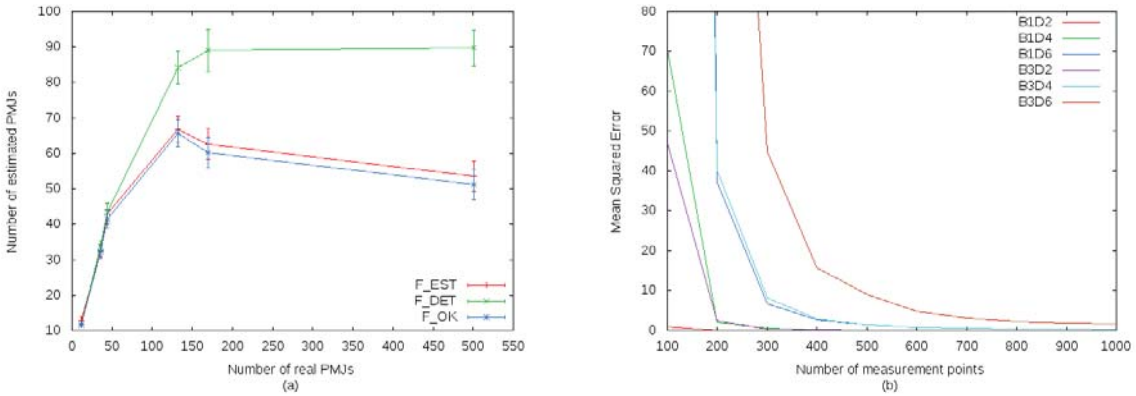


Figure 6: (a) Number of PMJs detected with respect to the total number of PMJs with 1000 measurement points. (b) MSE for the different scenarios considered.

in a cardiac tissue. We show that given an appropriate ratio between the number of electrical PMJs and samples acquired, the system can locate the effective PMJs and determine their activation time, provided that they are not arranged in small clusters. In order to detect all PMJs, we will need to sample, at least, with a density three times higher than the highest density of PMJs. When clustered electrical PMJs appear, they are virtually impossible to be detected with the current setup. However, PMJs within clusters have none or a very local impact in the overall activation map, since most of them are non-effective nodes. Overall mean time square errors obtained from the estimated PMJs are low in all the scenarios (including clustered) for EAMs with more than 600 samples acquired, which is a feasible number in clinics. In conclusion, we could reproduce the activations maps in a computer model using as an input the location and activation time of estimated PMJs.

The system does not recover the Purkinje tree structure of the system, and therefore it cannot be used in pathological scenarios in which PMJs are activated retrogradely, i.e. from tissue to the Purkinje tree. However, the result of our algorithm can be used as an input to Purkinje tree estimation methods. In addition, the anisotropic properties of the underlying tissue where the PMJs are placed has not been taken into account and should be included in future works. Other main future research directions are building a robust version of the algorithm, capable of estimating the PMJs from noisy LAT data and the extension of the domain to arbitrary geometries.

Acknowledgements

This research is funded by grants TIN2014-59932-JIN and TIN2015-66972-C5-5-R.

References

[1] A. Atkinson, S. Inada, J. Li, J. O. Tellez, J. Yanni, R. Sleiman, E. A. Allah, R. H. Anderson, H. Zhang, M. R. Boyett, and H. Dobrzynski, *Anatomical and molecular mapping of the left and right ventricular His-Purkinje conduction networks*, J Mol Cell Cardiol, 51 (2011), pp. 689–701.

[2] E. Behradfar, A. Nygren, and E. J. Vigmond, *The role of purkinje-myocardial coupling during ventricular arrhythmia: A modeling study*, PLoS ONE, 9 (2014), p. e88000.

[3] P. M. Boyle, S. Mass, K. Nanthakumar, and E. J. Vigmond, *Transmural ik(atp) heterogeneity as a determinant of activation rate gradient during early ventricular fibrillation: mechanistic insights from rabbit ventricular models.*, Heart Rhythm, 10 (2013), pp. 1710–1717.

[4] R. Cárdenes, R. Sebastian, D. Soto-Iglesias, A. Berruezo, and O. Camara, *Estimation of Purkinje trees from electro-anatomical mapping of the left ventricle using minimal cost geodesics.*, Medical image analysis, 24 (2015), pp. 52–62.

[5] J. E. Dennis Jr and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, vol. 16, SIAM, 1996.

[6] A. Lopez-Perez, R. Sebastian, and J. M. Ferrero, *Three-dimensional cardiac computational modelling: methods, features and applications.*, Biomedical engineering online, 14 (2015), p. 35.

[7] S. Palamara, C. Vergara, D. Catanzariti, E. Faggiano, C. Pangrazzi, M. Centonze, F. Nobile, M. Maines, and A. Quarteroni, *Computational generation of the purkinje network driven by clinical measurements: The case of pathological propagations*, Int. J. Numer. Meth. Biomed. Engng, 30 (2014), pp. 1558–1577.

[8] R. Sebastian, V. Zimmerman, D. Romero, D. Sanchez-Quintana, and A. F. Frangi, *Characterization and modeling of the peripheral cardiac conduction system.*, IEEE Trans Med Imaging, 32 (2013), pp. 45–55.

[9] C. Vergara, S. Palamara, D. Catanzariti, F. Nobile, E. Faggiano, C. Pangrazzi, M. Centonze, M. Maines, A. Quarteroni, and G. Vergara, *Patient-specific generation of the purkinje network driven by clinical measurements of a normal propagation*, Med Biol Eng Comput., 52 (2014), pp. 813–26.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Parameter Extraction in Electron Devices by means of Polynomial Pattern Analysis

D. Barrera[1], M. J. Ibáñez[1], A. M. Roldán[2], J. B. Roldán[2] and R. Yáñez[1]

[1] *Department of Applied Mathematics, University of Granada, 18071-Granada, Spain*

[2] *Department of Electronics, University of Granada, 18071-Granada, Spain*

emails: dbarrera@ugr.es, mibanez@ugr.es, amroldan@ugr.es, jroldan@ugr.es, ryanez@ugr.es

Abstract

The determination of polynomial patterns in scattered data is a problem that appears in different branches of science and engineering, among them the determination of parameters in compact models of electron device [1], [2], [3]. In the latter case, we have deepen on the use of this technique to determine the number of straight line portions contained in a curve in an automatic manner. The algorithm developed, based on discrete orthogonal polynomials, can be used for parameter extraction purposes in the context of electron device compact modeling. It is based upon the isolation of straight line sections in experimental or simulated data and the determination of the slope of those curve sections to calculate one or more parameters within a compact modeling context. This technique allows also the identification of curve portions that correspond to polynomials of different degree, a tool that could be very useful in parameter determination of device models.

Key words: MOSFET, parameter extraction, polynomial patter finding
MSC 2000: 41A05, 41A15, 65D05, 65D07

1 Discrete Orthogonal Chebyshev Polynomials

In these section we will briefly describe basic properties of discrete Chebyshev polynomials that will be used in this work.

The discrete Chebyshev orthonormal polynomials are defined as the following hypergeometric series [4, 5, 6]:

$$\tilde{t}_n(x;L) = \frac{1}{n!}\sqrt{\frac{(2n+1)(L-n-1)!}{(L+n)!}}\sum_{k=0}^{n}(-1)^k\binom{n}{k}(x-k-L+1)_n(x-k+1)_n,$$

where $L \in \mathbb{N}$ is a parameter, $x \in [0, L-1]$ is the variable, and $n \in \mathbb{N}$, $0 \leq n \leq L-1$, is the degree of the polynomial.

They verify the three term recurrence relation

$$\tilde{t}_{n+1}(x;L) = \frac{1}{\alpha_n}(x - \frac{L-1}{2})\tilde{t}_n(x;L) - \frac{\alpha_{n-1}}{\alpha_n}\tilde{t}_{n-1}(x;L),$$

with

$$\alpha_n = \frac{n+1}{2}\sqrt{\frac{L^2-(n+1)^2}{(2n+1)(2n+3)}},$$

and they are orthogonal with respect to the scalar product $\langle f, g\rangle := \sum_{x_i=0}^{L-1} f(x_i)g(x_i)$.

2 Numerical Procedure

In this section we will explain the procedure followed to isolate de straight line portions in the data (experimental or simulated).

Let us consider a set of discrete data $\{x_i, y_i\}$, $i = 1, \ldots N$, with x_i equispaced with steplength h. We try to determine subsets of data $\{x_i, y_i\}$, $i = i_m, \ldots, i_M$ that form straight lines.

The procedure can be divided in two parts: Given a subset of data, i) determine if these data form a straight line, and ii) select the subsets. We analyze these two facets below.

Let us take a data subset $\{x_i, y_i\}$, $i = i_1, \ldots, i_M$ and consider the scalar products,

$$r_n = \langle y, \tilde{t}_n(\cdot, M)\rangle = \sum_{i=0}^{M-1} y_i\tilde{t}_n(i, M).$$

The data can be assumed to be a linear combination of the polynomials $\tilde{t}_n$:

$$y_i = \sum_{j=0}^{\infty} c_j^i\tilde{t}_j(x_i, M),$$

so the scalar product r_n takes the value $r_n = c_n$.

We have made use of the orthonormality of the polynomials t_n. So, if data y_i form a straight line, all the scalar products r_n will be 0, except r_0 and/or r_1. In practice, we will

consider only the scalar products r_n for $n = 0, \ldots, n_{\max}$, with $n_{\max}$ given. Then, we check if all the coefficients r_n, $n = 0, \ldots, n_{\max}$, vanish except r_0 and/or r_1. Only in this case, we will consider the data to be a straight line. In practice we will consider that a coefficient r_n vanish if $|r_n| < \epsilon$. Of course, it can happens that data y_i come from a combination of t_0, t_1 and other polynomials with degrees higher than $n_{\max}$; nevertheless, considering that we can choose $n_{\max}$ sufficiently large and that our data are experimental and, so, affected by measuring errors, this is highly unlikely.

Once we know how to check if a given set of data forms a straight line, we have to determine the straight line subsets of data in the whole data set. To do so, (a) we start considering a subset of data with, at least, 4 points, with $i_1 = 1$ and $i_M = 4$, and check if this minimum set of points forms a straight line; (b) if it is so, we begin a binary search for larger subsets of straight line data between $i_M = 4$ and N (the total number of data points), ending when it detects the maximum subset of straight line data points; (c) once we have found the maximum subset of straight line data points, we restart the procedure by setting the first point to check to the next point $i_i = i_M + 1$ (the last point of the previous straight line plus 1) and $i_M = i_1 + 4$; (d) if the initial minima data points (i_1 and $i_1 + 4$) do not form a straight line, we increase each by 1 ($i_1 \to i_1 + 1$ and $i_M \to i_M + 1$) and start the procedure again; and (e) we stop the whole process when there are not enough data to consider a straigh line.

Once we have the subsets of straight line data, we compute the straight lines themselves, as $p(x) = c_0 t_0(x, M) + c_1 t_1(x, M)$, by polynomial interpolation or by polynomial fitting. This procedure can be straightforwardly generalized to detect substes of data with polynomial degree higher than 1.

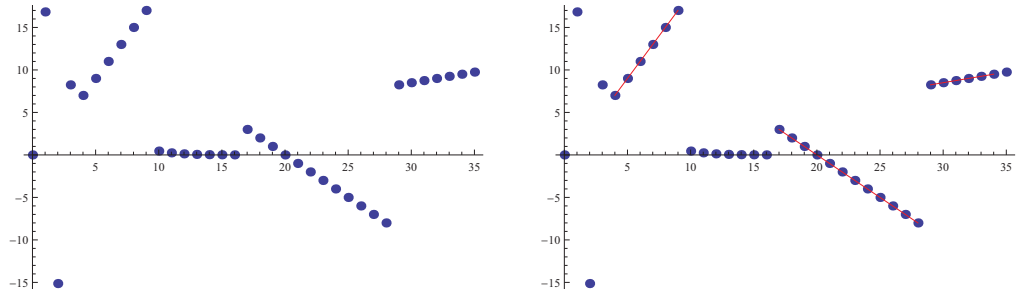
3 Practical Case

Here we will consider the function

$$f(x) = \begin{cases} 20\sin(x^2), & 0 \le x < 4, \\ 2x - 1, & 4 \le x < 10, \\ x^4 e^{-x}, & 10 \le x < 17, \\ -x + 20, & 17 \le x < 29, \\ x/4 + 1, & 30 \le x \le 35. \end{cases}$$

Next figure (left) shows the result of sampling $f(x)$ on the integers in the interval $[0, 35]$.

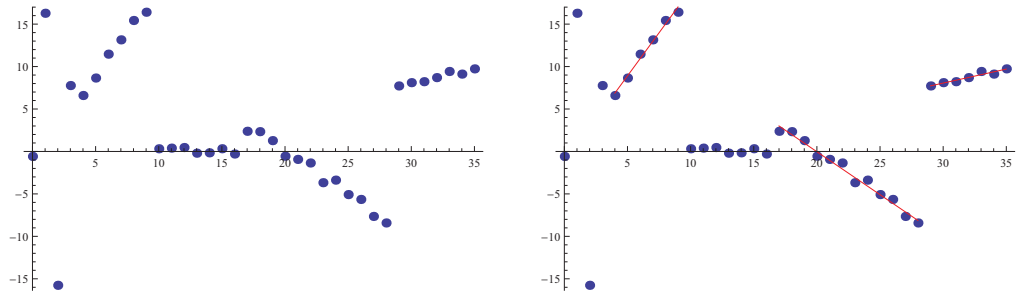
 ISBN: 978-84-608-6082-2



Clearly it shows 3 straight line sets, one for $x \in [4, 10[$, other for $x \in [17, 29[$ and the last one for $x \in [30, 35]$. On the naked eye, the points for $x \in [10, 17]$ can be mistaken as a straight line or a parabolic curve. The aplication of the previous algorithm provides the result shown in the previous figure (right).

As expected, data on $x \in [10, 17[$ are not detected as a straight line.

We modify the previous data randomly (simulating experimental data) to obtain the following figure (left)



In this case, we set the orthogonality threshold ϵ to 0.1, finding the previous straight data sets, as shown in the previous figure (right).

Again, the data on the interval $x \in [10, 17]$ are not detected as a stright line.

This technique can be used in the context of electron device compact modeling. There are many curves (current, capacitance, etc.) versus voltage or other magnitudes that show straight line segments. The finding of these straight lines in experimental data is employed to localize voltage values that are used as model parameters; in addition, the slope of these lines might be other parameters. That is why the use of an accurate method like the one presented here is of most important in this engineering context. For instance, in the context of determination of the threshold voltage in MOSFETs, the current versus gate-source voltage curve in logarithmic scale can be used to detect the linear region that corresponds to the subthreshold region, the end of this region can be defined as threshold voltage, and this can be considered a good approximation to this parameter.

References

[1] Arora, N.: MOSFET modelling for VLSI simulation. Theory and practice. World Scientific (2007)

[2] Ibáñez, M. J., Roldán, J. B., Roldán, A. M., Yáñez, R.: A comprehensive characterization of the threshold voltage extraction in MOSFETs transistors based on smoothing splines. Mathematics and Computers in Simulation, (102), 1–10 (2014)

[3] Gonzalez, P., Ibáñez, M. J., Roldán, A. M., Roldán, J. B.: An in-depth study on WENO-based techniques to improve parameter extraction procedures in MOSFET transistors. Mathematics and Computers in Simulation, (118), 248-257 (2015)

[4] Nikiforov, A. F., Uvarov, V. B.: Special Functions of Mathematical Physics. Birkhäuser Verlag, Basilea (1988)

[5] Nikiforov, A. F., Suslov, S. K., Uvarov, V. B.: Classical Orthogonal Polynomials of a Discrete Variable. Springer Series in Computational Physics. Springer-Verlag, Berlin (1991)

[6] Álvarez-Nodarse, R.: Polinomios hipergeométricos clásicos y q-polinomios. Monografías del Seminario Matemático "García de Galdeano", 26. Prensas Universitarias de Zaragoza, Zaragoza, (2003)

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Numerical solution of Love's integral equation by quasi-interpolation

D. Barrera[1], F. Elmokhtari[2], M. J. Ibáñez[1] and D. Sbibih[2]

[1] *Department of Applied Mathematics, University of Granada, 18071-Granada, Spain*

[2] *URAC05, FSO, University Mohammed I, URAC05, FSO, Oujda, Morocco*

emails: dbarrera@ugr.es, felmokhtari@correo.ugr.es, elmokhtari.fadila@gmail.com, mibanez@ugr.es, sbibih@yahoo.fr

Abstract

A fourth order convergent method is proposed to approximate the solutions to Love's integral equations. It is based on a particular approximation of the kernel by a C^1 quadratic quasi-interpolant. This results in a approximate integral equation whose solution is determined by solving a system of linear equations. All coefficients in this system can be exactly computed, and no quadrature formulae are needed.

Key words: Quadratic spline, Quasi-interpolant, Superconvergence, Love's integral equation
MSC 2000: 45L05, 65R20, 65D30

1 Introduction

In this paper, we describe a simple numerical method for solving the Love's integral equations given by

$$u(x) \pm \frac{1}{\pi} \int_{-1}^{1} \frac{d}{d^2 + (x-t)^2} u(t)\, dt = 1, \quad |x| \leq 1, \quad 0 \leq d \in \mathbb{R} \tag{1}$$

and which arise in the electrostatic problem of a circular plate condenser in an unbounded perfect fluid [8].

There is no closed expressions for the solutions of equations (1), so that several methods for approximating its solution have been proposed in the literature [1, 2, 3, 5, 6, 7, 9, 10, 11].

 ISBN: 978-84-608-6082-2

Here, we describe a simple and easy numerical method, based on C^1 quadratic splines, for approximating the solution of (1). It consists in approximating the univariate right section

$$t \to K(x,t) := \frac{1}{\pi} \frac{d}{d^2 + (x-t)^2}$$

of the kernel K by a C^1 quadratic spline quasi-interpolant.

We prove that this method has an order of convergence four.

2 Quadratic spline quasi-interpolant

Let $\tau_n := \{x_i, 0 \leq i \leq n\}$ be the uniform partition of the interval $I := [0,1]$ with meshlength $h = \frac{2}{n}$. We denote by $S_2(I, \tau_n)$ the space of splines of degree 2 and class C^1 on I. A basis of this space is formed by the $n+2$ quadratic B-splines $\{B_j, 0 \leq j \leq n+1\}$ associated with the extended partition of τ_n obtained by considering multiple knots at the endpoints, i.e. $x_{-2} = x_{-1} = x_0 = -1$ and $1 = x_n = x_{n+1} = x_{n+2}$. The support of B_j is equal to $[x_{j-2}, x_{j+1}]$.

We consider the classical discrete quasi-interpolant (see [12, 13])

$$Qf = \sum_{j=0}^{n+1} \mu_j(f) B_j,$$

where

$$\begin{aligned}
&\mu_0(f) = f_0,\ \mu_1(f) = \frac{1}{6}(-2f_0 + 9f_1 - f_2),\\
&\mu_j(f) = \frac{1}{8}(-f_{j-1} + 10f_j - f_{j+1}),\ 2 \leq j \leq n-1,\\
&\mu_n(f) = \frac{1}{6}(-f_{n-1} + 9f_n - 2f_{n+1}),\ \mu_{n+1}(f) = f_{n+1},
\end{aligned}$$

and $f_j := f(\theta_j)$ with $\theta_j := \frac{1}{2}(x_{j-1} + x_j)$.

Since the operator Q is exact on $\mathbb{P}_2$, for any function $f \in C^3(I)$ it holds (see [4])

$$\|f - Qf\|_{\infty,I} \leq C(1 + \|Q\|_\infty) h^3 \|f^{(3)}\|_{\infty,I},$$

where $\|.\|_{\infty,I}$ denotes the infinity norm on the interval I and C is a constant independent of h. Therefore,

$$\|f - Qf\|_\infty = \mathcal{O}(h^3).$$

It is useful to write the operator Q under the form

$$Qf = \sum_{j=0}^{n+1} f_j L_j,$$

where the functions L_j are defined as follows:

$$L_0 = B_0 - \frac{1}{3}B_1, \quad L_1 = \frac{3}{2}B_1 - \frac{1}{8}B_2, \quad L_2 = -\frac{1}{6}B_1 + \frac{5}{4}B_2 - \frac{1}{8}B_3,$$
$$L_j = -\frac{1}{8}B_{j-1} + \frac{5}{4}B_j - \frac{1}{8}B_{j+1}, \ 3 \leq j \leq n-2,$$
$$L_{n-1} = -\frac{1}{8}B_{n-2} + \frac{5}{4}B_{n-1} - \frac{1}{6}B_n, \quad L_n = -\frac{1}{8}B_{n-1} + \frac{3}{2}B_n, \quad L_{n+1} = -\frac{1}{3}B_n - B_{n+1}.$$

The operator Q provides the order of convergence $\mathcal{O}\left(h^3\right)$ on the whole interval I for regular enough functions. Moreover, it interpolates the monomial $m_3(x) := x^3$ at the knots x_i and θ_i, and, as a consequence the following result holds:

Proposition 1 *Assume that $f \in C^4(I)$, then we have:*

$$Qf(x_i) - f(x_i) = -\frac{3}{128}h^4 f^{(4)}(x_{i-1}) + \mathcal{O}\left(h^5\right), \qquad i = 2, \ldots, n-2,$$
$$Qf(\theta_i) - f(\theta_i) = -\frac{1}{64}h^4 f^{(4)}(\theta_{i-1}) + \mathcal{O}\left(h^5\right), \qquad i = 3, \ldots, n-2.$$

This superconvergence property implies that

$$|K(x, \theta_j) - K_n(x, \theta_j)| \leq \frac{1}{64}h^4 \left\|D^{0,4}K\right\|_{\infty, I^2}.$$

3 Right approximation of the kernel by quasi-interpolation

The approximation of the kernel K with respect its second variable by the above quasi-interpolant Q leads to a degenerate kernel in the space $S_2(I, \tau_n)$. More precisely, this approximation can be written as

$$K(x,t) \approx K_n(x,t) := QK(x,\cdot)(t) = \sum_{i=0}^{n+1} K_i(x) L_i(t),$$

where the rational functions $K_i := K(\cdot, \theta_i)$, $0 \leq j \leq n+1$, are given by

$$K_i(x) = \frac{1}{\pi}\frac{d}{d^2 + (x - \theta_i)^2}, \ -1 \leq x \leq 1.$$

Then, the approximate solution u_n satisfies the equation

$$u_n(x) = 1 \pm \int_{-1}^{1} K_n(x,t)\, u_n(t) = 1 \pm \sum_{i=0}^{n+1} K_i(x) \int_{-1}^{1} L_i(t)\, u_n(t)\, dt, \tag{2}$$

 ISBN: 978-84-608-6082-2

and it can be written in the form

$$u_n(x) = 1 \pm \sum_{j=0}^{n+1} X_j K_j(x).$$

Substituting in the integral equation (2), we get

$$\sum_{j=0}^{n+1} X_j K_j(x) = \sum_{j=0}^{n+1} K_j(x) \int_{-1}^{1} L_j(t) \left(1 \pm \sum_{j=0}^{n+1} X_j K_j(t)\right) dt.$$

It is easy to verify that that the functions K_j are linearly independent. Then,

$$X_j = b_j \pm \sum_{i=0}^{n+1} A_{i,j} X_i, \quad 0 \le j \le n+1.$$

This linear system can be written as

$$(I \ \pm \ A) X = b$$

where $X := (X_0, X_1, \ldots, X_{n+1})^T$, and the coefficients of the matrix A and the vector b are respectively defined by

$$A_{i,j} := \int_{-1}^{1} K_i(t) L_j(t)\, dt \qquad \text{and} \quad b_j := \int_{-1}^{1} L_j(t)\, dt, \quad 0 \le i, j \le n+1.$$

All the above integrals are computed exactly.

In order to derive the convergence order, let $\mathcal{K}$ (resp. $\mathcal{K}_n$) be the integral operator with kernel K (resp. K_n), i.e.

$$\mathcal{K}u(x) := \int_{-1}^{1} K(x,t)\, u(t)\, dt, \quad \mathcal{K}_n u(x) := \int_{-1}^{1} K_n(x,t)\, u(t)\, dt, \quad x \in [-1,1].$$

Proposition 2 *It holds $\|\mathcal{K} - \mathcal{K}_n\|_\infty = \mathcal{O}(h^4)$. Therefore, there exists a constant C independent on h such that*

$$\|u - u_n\|_{\infty,I} \le C h^4 \|u\|_{\infty.I}.$$

References

[1] M. Agida and A.S. Kumar, A Boubaker Polynomials Expansion Scheme Solution to Random Love's Equation in the Case of a Rational Kernel, Electronic Journal of Theoretical Physics 7 (24) (2010) 319–326.

[2] K. E. Atkinson, *The numerical solution of integral equations of the second kind*, Cambridge University Press 1997.

[3] L. M. Delves and J. L. Mohamed, *Computational Methods for Integral Equations*, Combridge University Press, Combridge, 1985.

[4] R. A. DeVore and G. G. Lorentz, *Constructive approximation*, Springer-Verlag, Berlin, 1993.

[5] D. Elliott, A Chebyshev series method for the numerical solution of Fredholm integral equations, The Computer Journal 6 (1963) 102–111.

[6] A. S. Kumar, An analytical solution to applied mathematics-related Love's equation using the Boubaker polynomials expansion scheme, Journal of the Franklin Institute 347 (2010) 1755–1761.

[7] P. Lomthong, P. Huabsomboon and H. Kaneko, A modified Taylor series method to the classical Love's equation, in: Recent Researches in Mathematical Methods in Electrical Engineering and Computer Science, Proceedings of the 13th IASME/WSEAS International Conference on Mathematical Methods and Computational Techniques in Electrical Engineering (MMACTEE '11), G. Thomas, C. Fleaurant, T. Panagopoulos, E. Chevassus-Lozza (eds.), WSEAS Press, pp. 68–72.

[8] E. R. Love, The potential due to a circular parallel plate condenser, Mathematika 37 (1990) 217–231.

[9] G. V. Milovanović and D. Joksimović, Properties of Boubaker polynomials and an application to Love's integral equation, Appl. Math. Comput. 224 (2013) 74–87.

[10] P. Pastore, The numerical treatment of Love's integral equation having very small parameter, J. Comput. Appl. Math. 236 (2011) 1267–1281.

[11] Y. Ren, B. Zhang and H. Qiao, A simple Taylor-series expansion method for a class of second kind integral equations, J. Comput. Appl. Math. 110 (1999) 15–24.

[12] P. Sablonnière, Quadratic spline quasi-interpolants on bounded domains of R^d, $d = 1, 2, 3$. Spline and radial functions, Rend. Sem. Mat. Univ. Pol. Torino 61 (2003) 61–78.

[13] P. Sablonnière, On some multivariate quadratic spline quasi-interpolants on bounded domains. In Modern Developments in Multivariate Approximation, W. Haussmann et al. (eds), ISNM Vol. 145, Birkhauser Verlag (2003), pp. 263–278.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Basic reproduction number in a spatially structured model for gut microbiota

Carles Barril[1], Àngel Calsina[1] and Jordi Ripoll[2]

[1] *Departament de Matemàtiques, Universitat Autònoma de Barcelona*

[2] *Departament d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona*

emails: carlesbarril@mat.uab.cat, acalsina@mat.uab.cat, jripoll@imae.udg.edu

Abstract

An ecological model for a bacterial population inside and outside an animal host is studied. We consider a linearised system for the proliferating bacteria, where they can be in three compartments: attached to the epithelial wall of the intestine or as free particles in the lumen or in the outer environment. The geometry of the intestine is reduced to a line segment where spatial densities are taken into account. We compute the next-generation operator as the composition of the cell division operator and the inverse of the transition/mortality operator. The basic reproduction number is then explicitly computed as the spectral radius of this linear operator. In addition, the extinction threshold is interpreted in terms of the expected number of bacteria coming back to the outer medium from each initial single bacterium, after travelling along the intestine. Further developments can include several hosts and/or the interaction with a bacteriophage population, and the analysis of evolutionary aspects of the model.

Key words: bacterial populations, next-generation operator

1 Introduction

We study the interaction between bacteria and bacteriophage populations inhabiting both the intestine of an animal host and the external environment, see [1] and the references therein, and [2]. We are interested in the asymptotic behaviour of the proliferating bacteria. Our aim in the present work is to determine their final outcome, either extinction or non-limited growth, for the case corresponding to the absence of phages and assuming Malthusian growth of the bacteria.

2 The Model

Let us consider a population of bacteria either proliferating along the intestine of an animal host or living outside in the environment. Assuming radial homogeneity in space, the intestine of an animal can be described by a line segment of length $l > 0$ where a flow of constant velocity $c > 0$ is assumed. Let x be space (intestine starts at $x = 0$ and ends at $x = l$) and let $t \geq 0$ be time. The state variables are: $u(x,t)$, the density with respect to space of bacteria attached to the intestinal epithelium, $v(x,t)$, the density with respect to space of bacteria in the intestinal lumen, and $b(t)$, the total population of bacteria in the exterior medium adjacent to the host.

Let us assume that the bacteria populations within the host grow at constant per capita rates γ_1 and γ_2 depending on whether they are attached or free. Let us denote the per capita rates of *attachment*, *detachment*, *reinfection* (recruitment from the environment to the host) and *mortality* at the exterior medium by $\alpha, \delta, \rho, \mu$ respectively. Then, the spatially structured ecological model reads as:

$$\begin{cases} \partial_t u(x,t) = \gamma_1 u(x,t) + \alpha v(x,t) - \delta u(x,t) \\ \partial_t v(x,t) = -c\partial_x v(x,t) + \gamma_2 v(x,t) - \alpha v(x,t) + \delta u(x,t) \\ b'(t) = cv(l,t) - \rho b(t) - \mu b(t) \\ cv(0,t) = \rho b(t) \end{cases} \tag{1}$$

Growth rates here take into account both cell division and cell mortality. Although the computations below can be done analogously, for the sake of simplicity, we will assume that $\gamma_1 \geq 0$ and $\gamma_2 \geq 0$ only correspond to cell division rates.

Notice that linear system (1) is a combination of PDEs and ODEs and the solution of the system defines a strongly continuous semigroup of positive bounded linear operators on the Banach lattice $X = L^1(0,l) \times L^1(0,l) \times \mathbb{R}$. The standard approach to the asymptotic behaviour of (1) based on computing the spectral bound (*intrinsic growth rate*) of the infinitesimal generator of the semigroup yields to an implicit expression for the growth rate of the population via a characteristic equation. Instead, here we take the approach of the next-generation operator which is equivalent and gives an explicit expression for the basic reproduction number. Moreover, additional evolutionary aspects can be analysed once this expression is obtained, according to different biologically meaningful parameters of the model.

3 Next-generation operator

In this section we compute and analyze the next-generation operator, see [3], for the linear system of bacterial population stated above. Next-generation operator is related to the cell

division operator (a cell gives rise to two daughter cells) defined on X as:

$$B(u,v,b) = (2\gamma_1 u, 2\gamma_2 v, 0)$$

and to the transition/mortality operator (which takes into account that cells disappear when they divide), defined on X as:

$$M(u,v,b) = (\gamma_1 u - \alpha v + \delta u, cv' + \gamma_2 v + \alpha v - \delta u, -cv(l) + \rho b + \mu b),$$

with domain $D = L^1(0,l) \times \{(v,b) \in W^{1,1}(0,l) \times \mathbb{R} : cv(0) = \rho b\}$ which is a dense subspace of X. More precisely, it is computed as BM^{-1}, the composition of the cell division operator and the inverse of the transition/mortality operator:

$$\begin{pmatrix} 2\gamma_1 & 0 & 0 \\ 0 & 2\gamma_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \gamma_1 + \delta & -\alpha & 0 \\ -\delta & c\partial_x + \gamma_2 + \alpha & 0 \\ 0 & -c\delta_l & \rho + \mu \end{pmatrix}^{-1}$$

which turns out to be a positive bounded linear operator on X . Here δ_l stands for the Dirac delta operator at the point $x = l$.

In order to avoid degenerate cases, we will assume in the sequel that $\alpha, \delta, \rho > 0$, $\mu \geq 0$, and $\gamma_1 + \gamma_2 > 0$ (i.e. that they are not both zero).

3.1 Basic reproduction number

We are going to compute the basic reproduction number $\mathcal{R}$ which is defined as the expected number of newborns produced by each bacterium during its lifetime (in the long run). This number is computed as $\mathcal{R} = r(BM^{-1})$, the spectral radius of the next-generation operator, see [3].

First of all, let us recall that the spectral radius of a positive bounded linear operator always belongs to the spectrum of the operator, see e.g. [4]. So the spectral radius of BM^{-1} is actually an spectral value of BM^{-1}.

Let us consider the operator $L_\lambda \hat{\phi} := BM^{-1}\hat{\phi} - \lambda\hat{\phi}$ defined on X. Denoting $\phi = M^{-1}\hat{\phi}$, we have that $M\phi = \hat{\phi}$ and hence that L_λ is surjective if and only if Range$(B - \lambda M)$ is the whole space X. Notice that $\tilde{\lambda} = \frac{2\gamma_1}{\gamma_1 + \delta}$ belongs to the spectrum $\sigma(BM^{-1})$. Indeed, the first component of the image by $(B - \tilde{\lambda}M)$ of any $(u,v,b) \in D$ equals to $\tilde{\lambda}\alpha v \in W^{1,1}(0,l)$. Hence, $(B - \tilde{\lambda}M)$ is not surjective. Notice that $0 \leq \tilde{\lambda} < 2$.

In fact, it can be shown that BM^{-1} always has a unique real spectral value larger than the value $\tilde{\lambda}$ computed above, and therefore the former will coincide with the spectral radius. Moreover, it is an eigenvalue. Since such an eigenvalue is precisely the basic reproduction number, we use the symbol $\mathcal{R}$ to denote it. By means of a characteristic equation, an explicit formula for $\mathcal{R}$ is obtained. Specifically,

$$\mathcal{R} = \frac{2}{1 - z^*}$$

 ISBN: 978-84-608-6082-2

where z^* is the unique solution of the equation

$$F(z) := \frac{\rho}{\rho+\mu} e^{-\left(\frac{\alpha\gamma_1 z}{\gamma_1 z+\delta}+z\gamma_2\right)\frac{l}{c}} = 1 \, .$$

Since the function $F(z)$, defined for $z \in (-\delta/\gamma_1, 1]$, decreases, tends to infinity as $z \downarrow -\frac{\delta}{\gamma_1}$ and fulfils $F(0) = \frac{\rho}{\rho+\mu} \leq 1$, the existence of a unique $z^* \in (-\frac{\delta}{\gamma_1}, 0]$ such that $F(z^*) = 1$ is guaranteed. Moreover, $F(z^*) = 1$ is transformed into $\left(\frac{\alpha\gamma_1 z^*}{\gamma_1 z^*+\delta} + z^*\gamma_2\right) = \frac{c}{l} \ln \frac{\rho}{\rho+\mu} =: A \leq 0$ and thus

$$z^* = \frac{-(\alpha\gamma_1 + \delta\gamma_2 - \gamma_1 A) + \sqrt{(\alpha\gamma_1 + \delta\gamma_2 - \gamma_1 A)^2 + 4\gamma_1\gamma_2\delta A}}{2\gamma_1\gamma_2} \, .$$

As expected, the basic reproduction number cannot be bigger than 2 since $z^* \leq 0$ and so $\mathcal{R} = \frac{2}{1-z^*} \leq 2$. Summarizing, $\mathcal{R} > 1$, i.e. there is non-limited growth of the bacterial population, see [5], when $\gamma_1 \geq \delta$ (since then $\mathcal{R} > \tilde{\lambda} \geq 1$) or when $\gamma_1 < \delta$ and $F(-1) = \frac{\rho}{\rho+\mu} e^{\left(\frac{\alpha\gamma_1}{\delta-\gamma_1}+\gamma_2\right)\frac{l}{c}} > 1$ (which amounts to $z^* > -1$ and so to $\mathcal{R} > 1$). On the contrary, $\mathcal{R} < 1$, i.e. we have extinction of the bacterial population, when $\gamma_1 < \delta$ and $\frac{\rho}{\rho+\mu} e^{\left(\frac{\alpha\gamma_1}{\delta-\gamma_1}+\gamma_2\right)\frac{l}{c}} < 1$.

3.2 Biological interpretation of the extinction threshold

For the case $\gamma_1 < \delta$, the key quantity for the extinction threshold $\mathcal{R} = 1$ is

$$\frac{\rho}{\rho+\mu} e^{\left(\frac{\alpha\gamma_1}{\delta-\gamma_1}+\gamma_2\right)\frac{l}{c}} \, . \tag{2}$$

It can be interpreted as the expected number of bacteria leaving the intestine for the first time and that descend from an initial bacterium in the external medium. In order to show this result one can take a model similar to (1) but assuming now that the bacteria leaving the intestine are removed from the system instead of going back to the external media. The initial condition of having a founder bacterium in the external media is equivalent to saying that, with probability $\rho/(\rho+\mu)$, the bacterium is at the beginning of the intestine (otherwise it has died before being ingested by the host). This leads us to the initial value problem

$$\begin{cases} \partial_t u(x,t) = \gamma_1 u(x,t) + \alpha v(x,t) - \delta u(x,t) \\ \partial_t v(x,t) = -c\partial_x v(x,t) + \gamma_2 v(x,t) - \alpha v(x,t) + \delta u(x,t) \\ cv(0,t) = 0 \\ v(x,0) = \delta_0(x) \end{cases} \, . \tag{3}$$

Note that $cv(l,t)$ is the expected density of bacteria that are leaving the intestine (hence the system) at time t. Using some previous results due to [2], it is shown that if $\gamma_1 \geq \delta$ then

$$\int_0^\infty cv(l,t)dt = \infty$$

whereas if $\gamma_1 < \delta$ then

$$\int_0^\infty cv(l,t)dt = e^{\left(\frac{\alpha\gamma_1}{\delta-\gamma_1}+\gamma_2\right)\frac{l}{c}}.$$

Another way to get the same results about the interpretation of the quantity (2) is by means of a model that takes into account a constant supply $\beta > 0$ of bacteria to the external media, namely

$$\begin{cases} \partial_t u(x,t) = \gamma_1 u(x,t) + \alpha v(x,t) - \delta u(x,t) \\ \partial_t v(x,t) = -c\partial_x v(x,t) + \gamma_2 v(x,t) - \alpha v(x,t) + \delta u(x,t) \\ b'(t) = -\rho b(t) - \mu b(t) + \beta \\ cv(0,t) = \rho b(t) \end{cases}. \tag{4}$$

It is easy to show that the solutions of this system converge to an equilibrium $(\bar{u}, \bar{v}, \bar{b})$ provided $\gamma_1 < \delta$. In this case, the expected number of bacteria that leave the intestine for the first time and that descend from an initial bacterium in the external medium can be expressed as the ratio

$$\frac{c\bar{v}(l)}{\beta}$$

which turns out to be the quantity $\frac{\rho}{\rho+\mu}e^{\left(\frac{\alpha\gamma_1}{\delta-\gamma_1}+\gamma_2\right)\frac{l}{c}}$. If $\gamma_1 \geq \delta$, then the bacterial population grows forever and in some sense the above ratio becomes infinite.

Acknowledgements

This work has been partially supported by the coordinated projects MTM2014-52402-C3-2 and MTM2014-52402-C3-3 of the Spanish government. JR is also partially supported by the project MPCUdG2016/047 of the University of Girona and CB by the Spanish Ministry of Education grant FPU13/04333.

References

[1] C. Barril, A. Calsina, *Stability analysis of an enteropathogen population growing within a heterogeneous group of animals*, UAB Prepub. 15, Nov. 2015, 1–19.

[2] B. Boldin, *Persistence and spread of gastro-intestinal infections: the case of enterotoxigenic Escherichia coli in piglets*, Bull. Math. Biol., **70**, (2008), 7, 2077–2101.

[3] O. Diekmann, J.A.P. Heesterbeek, J.A.J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol., **28**, (1990), 4, 365–382.

[4] H.H. Schaefer, *Banach lattices and positive operators*, Springer, 1974.

[5] H. Thieme, *Spectral bound and reproduction number for infinite-dimensional population structure and time heterogeneity*, SIAM J. Appl. Math., **70**, (2009), 1, 188–211.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A family of second derivative free forth order continuation method for solving nonlinear equations

R.Behl[1], P.Maroju[1] and S.S.Motsa[1]

[1] *Department of Mathematics, Statistics and Computer science, University of KwaZulu-Natal,Private Bag X01, Scottsville 3209, Pietermaritzburg, South Africa*

emails: ramanbehl87@yahoo.in, maroju.prashanth@gmail.com, sandilemotsa@gmail.com

Abstract

We investigate a parametrized set of iterative method free from second derivative for solving nonlinear equations of the type $f(x) = 0$. This method can be thought of as a weighted mean between the Chebyshev and the Halley methods, the weights being α and $1 - \alpha$, where $\alpha \in \mathbb{R}$. Ezquerro and Hernandez[3] developed the uniparametric family of iterative method and convergence analysis in Banach space. Based in his idea, First we write the method in $\mathbb{R}$ then, we developed uniparametric family of second derivative free iterative method in $\mathbb{R}$. The convergence orders of the proposed method is discussed for different value of parameter α. We observed that for $\alpha = 2$, the order of iterative method is four and remaining values of parameter $\alpha \in \mathbb{R}$ order three. Several numerical examples are given to illustrate the efficiency and performance of the method for different values of α. Finally, we observed that Halley's method is more suitable for solving nonlinear equations than Chebyshev method. Even more, we can consider other iterations for α more suitable than Halley's and Chebyshev method. Also, we have compared our method with the Halley's method(HM) and the Chebyshev method (CM) by basins of attraction and observed that the proposed scheme is more efficient.

Key words: A continuation method, Nonlinear equations, The Halley's method, The Chebyshev's method

1 Introduction

Solving non-linear equations is one of the most important and challenging problems in numerical analysis. Let $f : D \subseteq \mathbb{R} \to \mathbb{R}$ be a nonlinear differentiable on a open interval D.

One of the main problems in numerical analysis is to solve the nonlinear equation

$$f(x) = 0 \tag{1}$$

Newton's method is basic and well known iterative method for solving nonlinear (1). The order of convergence of this method is quadratically convergent. Main advantage of this method is that the computation of second derivative not required. Also, we have many higher order iterative methods for solving nonlinear equations. The well know third order iterative methods for solving nonlinear equation (1) are the Chebyshev's method, the Halley's method and the Super-Halley's method [12, 6]. Chebyshev-Halley's method is well known parametric iterative method for solving nonlionear equations. The order of convergence of this method is three. In recent years, many variants of Chebyshev-Halley's methods with higher order iterative methods studied by many authors [4, 7]. They use the approach that the computation of second derivative at another point. These methods are efficient for the case the second derivative cost inexpensively. However, these methods depend on computationally cost second derivative, so that practical applications restricted. So, we use Newton's method [9]which quadratically convergent is frequently used to solve non-linear equations because of higher computational efficiency. The convergence of second derivative free uniparametric family of iterative method discussed by [5, 1, 2].

A Continuation method is a well known parametric based iterative method for solving nonlinear equations. Prashanth and Gupta [10] studied the the convergence of second derivative free continuation method between the Chebyshev and the Super-Halley methods in $\mathbb{R}$. According to the basic idea of continuation methods, a homotopy $\alpha h_1(x)+(1-\alpha)h_2(x)$, where $\alpha \in [0,1]$, can be defined between two functions h_1 and h_2. Based on this idea, Ezquerro and Hernandez[3] designed the uniparametric family of iteration method in Banach space. Motivated by this, we write the uniparametric family of iteration method for solving nonlinear equations (1),

$$x_{\alpha,n+1} = x_{\alpha,n} - (1 + \frac{1}{2}L_f(x_{\alpha,n})(1 + \frac{\frac{\alpha}{2}L_f(x_{\alpha,n})}{1 - \frac{1}{2}L_f(x_{\alpha,n})})) \times \frac{f(x_{\alpha,n})}{f'(x_{\alpha,n})}, \tag{2}$$

where,

$$L_f(x_{\alpha,n}) = \frac{f(x_{\alpha,n})f''(x_{\alpha,n})}{f'(x_{\alpha,n})^2} \tag{3}$$

However, this method depends on the second derivatives in computing process and its practical application is restricted rigorously. For this reason, In this paper, we develop the new modified variant of uniparametric family of continuation method in $\mathbb{R}$. We discuss the convergence analysis of these variants. The analysis shows that order of convergence of this method is three for parameter $\alpha \in \mathbb{R}$ and four for $\alpha = 2$. This method require three

evaluations of the function and one of its derivative. Some numerical examples are worked out to show the efficiency and superiority of the new method.

The rest of the paper is organized as follows. We developed the second derivative free continuation method and the convergence analysis of the method discussed in Section 2. In Section 3, we worked out some numerical examples to describe the performance of our method. The basin of attractions compare with existing methods discussed in Section 3. Finally, the conclusion given in Section 4.

2 The Method and Analysis of convergence

Let us consider the sufficiently differentiable function f, using the Taylor's series, we have

$$\begin{aligned} f(y_{\alpha,n}) &\approx f(x_{\alpha,n}) + (y_{\alpha,n} - x_{\alpha,n})f'(x_{\alpha,n}) + \frac{(y_{\alpha,n} - x_{\alpha,n})^2}{2} f''(x_{\alpha,n}) \\ &= \frac{1}{2}\Big[\frac{f(x_{\alpha,n})}{f'(x_{\alpha,n})}\Big]^2 f''(x_{\alpha,n}) \end{aligned} \tag{4}$$

where, $y_{\alpha,n} = x_{\alpha,n} - f(x_{\alpha,n})/f'(x_{\alpha,n})$. From (4), we get

$$f''(x_{\alpha,n}) \approx \frac{2f(y_{\alpha,n})f'(x_{\alpha,n})^2}{f(x_{\alpha,n})^2}. \tag{5}$$

Replacing (5) in (3), we get

$$L_f(x_{\alpha,n}) = \frac{2f(y_{\alpha,n})}{f(x_{\alpha,n})} \tag{6}$$

Hence, we obtain second derivative free variant of uniparametric iterative method by using (6) in (2) for $x_{\alpha,n} \in \mathbb{R}$, $n = 0, 1, 2, \ldots$,

$$\left.\begin{aligned} y_{\alpha,n} &= x_{\alpha,n} - \tfrac{f(x_{\alpha,n})}{f'(x_{\alpha,n})} \\ x_{\alpha,n+1} &= y_{\alpha,n} - \left(\tfrac{f(y_{\alpha,n})}{f(x_{\alpha,n})} \times \tfrac{(f(x_{\alpha,n})+(\alpha-1)f(y_{\alpha,n}))}{(f(x_{\alpha,n})-f(y_{\alpha,n}))}\right)\tfrac{f(x_{\alpha,n})}{f'(x_{\alpha,n})} \end{aligned}\right\} \tag{7}$$

Theorem 1 *Let the function $f : D \subseteq \mathbb{R} \to \mathbb{R}$ is twice differentiable and it has a simple root $x^* \in D$, where D is a open interval. For initial approximation $x_{\alpha,0}$, $\alpha \in \mathbb{R}$ close to the solution x^*, the iterative method (7) has a forth order of convergence for $\alpha = 2$ and third order for $\alpha \in \mathbb{R}$. The error equations is*

$$e_{\alpha,n+1} = -(-2+\alpha)c_2^2 e_n^3 - c_2((9-6\alpha)c_2^2 + (-7+4\alpha)c_3)e_{\alpha,n}^4 + O(e_{\alpha,n}^5). \tag{8}$$

Proof: Let $e_{\alpha,n} = x_{\alpha,n} - x^*$. Using Taylor's series expansion, we have

$$f(x_{\alpha,n}) = f'(x^*)[e_{\alpha,n} + c_2 e_{\alpha,n}^2 + c_3 e_{\alpha,n}^3 + c_4 e_{\alpha,n}^4 + O(e_{\alpha,n}^5)] \tag{9}$$

where, $c_k = \frac{f^k(x^*)}{k!f'(x^*)}$ for $k \in \mathbb{N}$. Also, expanding $f'(x_{\alpha,n})$ using Taylor's series around x^* we get

$$f'(x_{\alpha,n}) = f'(x^*)[1 + 2c_2 e_{\alpha,n} + 3c_3 e_{\alpha,n}^2 + 4c_4 e_{\alpha,n}^3 + O(e_{\alpha,n}^4)]. \tag{10}$$

Furthermore, we have

$$\begin{aligned} \frac{f(x_{\alpha,n})}{f'(x_{\alpha,n})} &= [e_{\alpha,n} + c_2 e_{\alpha,n}^2 + c_3 e_{\alpha,n}^3 + c_4 e_{\alpha,n}^4 + O(e_{\alpha,n}^5)][1 + 2c_2 e_{\alpha,n} + 3c_3 e_{\alpha,n}^2 + 4c_4 e_{\alpha,n}^3 + O(e_{\alpha,n}^4)]^{-1} \\ &= e_{\alpha,n} - c_2 e_{\alpha,n}^2 + 2(c_2^2 - c_3)e_{\alpha,n}^3 + (7c_2c_3 - 4c_2^2 - 3c_4)e_{\alpha,n}^4 + O(e_{\alpha,n}^5). \end{aligned} \tag{11}$$

Using this we get,

$$\begin{aligned} y_{\alpha,n} &= x_{\alpha,n} - \frac{f(x_{\alpha,n})}{f'(x_{\alpha,n})} \\ &= c_2 e_{\alpha,n}^2 + 2(-c_2^2 + c_3)e_{\alpha,n}^3 + (-7c_2c_3 + 4c_2^3 + 3c_4)e_{\alpha,n}^4 + O(e_{\alpha,n}^5). \end{aligned}$$

Expanding $f(y_{\alpha,n})$ by Taylor's series around x^*, we get

$$f(y_{\alpha,n}) = f'(x^*)[c_2 e_{\alpha,n}^2 + 2(-c_2^2 + c_3)e_{\alpha,n}^3 + (-7c_2c_3 + 5c_2^3 + 3c_4)e_{\alpha,n}^4 + O(e_{\alpha,n}^5)]. \tag{12}$$

From (11) and (12), we get

$$\begin{aligned} \Big(1 + \frac{f(y_{\alpha,n})}{f(x_{\alpha,n})} \frac{(f(x_{\alpha,n}) + (\alpha-1)f(y_{\alpha,n}))}{(f(x_{\alpha,n}) - f(y_{\alpha,n}))}\Big) \frac{f(x_{\alpha,n})}{f'(x_{\alpha,n})} &= e_n + (\alpha-2)c_2^2 e_n^3 \\ &\quad + c_2\left((9-6\alpha)c_2^2 + (4\alpha-7)c_3\right)e_n^4 \end{aligned} \tag{13}$$

Using $e_{\alpha,n+1} = x_{\alpha,n+1} - x^*$, the error equation and (13) in (7), we get

$$e_{\alpha,n+1} = -(\alpha-2)c_2^2 e_n^3 - c_2\left((9-6\alpha)c_2^2 + (4\alpha-7)c_3\right)e_n^4 + O(e_{\alpha,n}^5). \tag{14}$$

This means that iterative method defined by (7) is forth order convergent for $\alpha = 2$ and cubically convergent for $\alpha \in \mathbb{R}$

3 Numerical Examples

Now we employ iterative method (7) to solve some non-linear equations and compare these methods with Chebyshevs method, Halley's method. The computational results are displayed in Table 1. From numerical examples, we can see that iterative method (7) is

obviously superior to Chebeshevs method, Halley's method because these new methods do not require the second derivatives. Furthermore, we show also that iterative method (7) for $\alpha = 2$ is better than Chebeshevs method, Halley's method. So we feel that these new methods have the great practical utility. The test functions and their roots upto 16th decimal places are given below. The computational results given in Table 1. From this, we compare the residual of the function $|f(x_n)|$, $\delta = |x_{\alpha,n} - x^*|$ and Computational order of convergence (COC). we observe that iterative method (7) is better than the Chebyshev method (CM) , the Halley's method (HM). So that, the method have practical utility.

Example 1 $f_1(x) = x^2 - e^x - 3x + 2$, $x_0 = 2.5$, $x^* = 0.25753028543986084$,

Example 2 $f_2(x) = x^2 \sin(x) - \cos(x)$, $x_0 = 6$, $x^* = 6.3083089552381511$,

Example 3 $f_3(x) = x^2 - (2 - x)^3$, $x_0 = 1.1$, $x^* = 1.0000000000000000$,

Example 4 $f_4(x) = (x + 2)e^{-x} + x$, $x_0 = -7$, $x^* = -1.6878939988284736$,

Table 2:Comparison of Number of Iterations, $|f(x_{\alpha,n})|$, $|x_{\alpha,n} - x^*|$, COC

Examples	CM	HM	Method (7)($\alpha = 2$)	Method (7)($\alpha = 3/2$)
f_1, $x_0 = 1$				
IT	5	5	4	5
$\|f(x_{\alpha,n})\|$	3.3e-72	1.2e-99	2.3e-93	9.4e-139
COC	3.0000	3.0000	4.0000	3.0000
δ	8.6e-738	3.3e-100	6.0e-94	2.5e-139
f_2, $x_0 = 6$				
IT	4	4	4	4
$\|f(x_{\alpha,n})\|$	2.1e-50	2.0e-61	5.8e-172	5.3e-98
COC	3.0000	3.0000	4.0000	3.0000
δ	5.1e-52	5.0e-63	1.4e-173	1.3e-99
f_3, $x_0 = 4.29$				
IT	4	4	3	3
$\|f(x_{\alpha,n})\|$	2.3e-123	2.7e-130	7.9e-85-	3.7e-126
COC	4.0000	3.0000	4.0000	3.0000
δ	4.6e-124	5.4e-131	1.6e-85	7.3e-127
f_4, $x_0 = -7$				
IT	10	5	8	8
$\|f(x_{\alpha,n})\|$	1.2e-41	8.6e-80	3.1e-38	1.4e-35
COC	2.0000	3.0000	3.9994	3.0000
δ	2.6e-42	1.8e-80	6.7e-38	2.9e-36

 ISBN: 978-84-608-6082-2

4 Attractor basins in the complex plane

The dynamical properties of rational functions give us important information about numerical features of the iterative method. We here investigate the comparison of the attained root finders in the complex plane using basins of attraction. To generate the basins of attraction for the zeros of a polynomial and an iterative method, we take grid of 400×400 points in a rectangle $D = [-3,\ 3] \times [-3,\ 3] \in \mathbb{C}$ and we use these points as z_0. If the sequence generated by the iterative method reaches to the solution z^* of the polynomial with the tolerance $|z_n - z^*| < 10^{-5}$ and a maximum of 25 iterations. If the initial point z_0 is the basin of attraction of the zero and we paint this point in a colour previously selected for the root.In the same basin of attraction, the number of iteration needed to acheive the solution is showed in darker or brighter colour. Black colour denotes lack of convergence to any of the roots or convergence to infinity. In this way, we distinguish the attraction basins by their colors for different methods.

For the better comparison, we consider for three polynomial functions.

Test problem 1. Let $p_1(z) = (z-1)^3 - 1$, having simple zeros $\{0.5 - 0.866025i,\ 0.5 + 0.866025i,\ 2.0\}$. It is straight forward to see from the Fig. 1 that our method (7) has larger and brighter basin of attraction as compared to CM, HM and Method for $\alpha = 3/2$.

Test Problem 2. Let $p_2(z) = z^2 - 1$, having simple zeros $\{-1.0,\ -i,\ i,\ 1,\}$. We observed from the Fig. 2 that our method (7) performed has larger and brighter basin of attraction as compared to CM, HM and Method for $\alpha = 3/2$.

Test Problem 3. Let $p_3(z) = (z^4-1)$, having simple zeros $\{-1,\ 0,\ 0.5-0.866025i,\ 0.5+0.866025i\}$. From the Fig. 3, it is clear that method (7) has larger and brighter basin of attraction as compared to the methods namely, CM, HM and Method for $\alpha = 3/2$.

Figure 1: The basins of attraction for CM, HM and Method (7) for $\alpha = 3/2, 2$, respectively in problem 1.

Figure 2: The basins of attraction for CM, HM and Method (7) for $\alpha = 3/2, 2$, respectively in problem 2.

Figure 3: The basins of attraction for CM, HM and Method (7) for $\alpha = 3/2, 2$, respectively in problem 3.

5 Conclusions

In this paper, we obtained the family of uniparametric second derivative free iterative method for solving nonlinear equation. We prove that the order of convergence of iterative method is four for parameter $\alpha = 2$ and three for $\alpha \in \mathbb{R}$. Some numerical examples are worked out. We compare Number of iterations, $|f(x_n)|$, δ and Computational order of convergence (COC) with the Cheyshev's method, The Halley's method, it was observed that they demonstrate at better behavior. From numerical examples, we show that these new methods have the great practical utility. Finally, we find the basins of attraction using our method and compare with the Chebyshev and the Halley method, we observed that the proposed scheme is more efficient.

Acknowledgements

This work has been supported by University of KwaZulu-Natal.

References

[1] J.A. Ezquerro, M.A. Hernandez, *A uniparametric Halley-type iteration with free second derivative*, Int.J. Pure Appl. Math. **6** (2003) 103–114.

[2] J.Ezquerro, M.A.Hernández, *On Halley-type iteration with free second derivative*, J. Comput. Appl. Math. **170** (2004) 455–459.

[3] J.Ezquerro, M.A.Hernández, *On a class of iteration containing the chebyshev and the Halley methods*, Publ. Math. Debrecen. **54** (1999) 403–415.

[4] M.Grau, J.L.Diaz-Barrero, *An improvement of the Euler-Chebyshev iterative method*, J. Math. Anal. Appl. **315** (2006) 1–7.

[5] J.Kou, Y.Li and W.Xiuhua, *A uniparametric Chebyshev-type method free from second derivatives*, J. Appl. Math. Compu. **179** (2006) 296–300.

[6] J.Kou, I.K. Argyros, *A note on the Halley method in Banach spaces*, J. Appl. Math. Compu. **58** (1993) 215–224.

[7] J.Kou,Y.Li, *A family of modified super-Halley methods with fourth-order convergence*, J. Appl. Math. Compu. **189** (2007) 336–370.

[8] B. Neta, M. Scot and C. Chun, *Basins of attraction for several methods to find simple roots of nonnlinear equations*, J. Appl. Math. Compu. **218** (2012) 10548-10556.

[9] A.M. Ostrowski, *Solution of Equations in Euclidean and Banach Space*,third ed., Academic Press, New York, 1973.

[10] M.Prashanth, D. K.Gupta, *A continuation method for Solving nonlinear equations in R*, Int. J. Comput. Sci. and Mathe. **5** (2014) 209–218.

[11] M. Scott, B. Neta, C. Chun, *Basins attractors for various methods*, Appl. Math. Mathe. **218** (2011) 2584–2599.

[12] J.F. Traub, *Iterative Methods for Solution of Equations*,Prentice-Hall,Englewood Cliffs, NJ, 1964.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Performance Evaluation of the Iteratively Reweighted Least Squares Algorithm (IRLS) on a Multi-core Platform

Jose A. Belloch[1], Carla Ramiro[2], Enrique S. Quintana-Ortí[1] and Antonio M. Vidal[3]

[1] *Depto. de Ingeniería y Ciencia de Computadores, Universitat Jaume I de Castelló, Spain*

[2] *Instituo Universitario de Investigación, de Robótica y Tecnologias de la Información, Universitat de València, Spain*

[3] *Depto. de Sistemas Informáticos y Computación, Universitat Politècnica de València, Spain*

emails: jbelloch@uji.es, carla.ramiro@uv.es, quintana@uji.es, avidal@dsic.upv.es

Abstract

The Iteratively Reweighted Least Squares algorithm (IRLS) has been applied to numerous optimization problems, such as the coefficient design of FIR and IIR filters, which are widely used in signal processing applications. The IRLS algorithm presents a high computational cost, especially when real-time conditions are demanded. The appearance of the multi-core platforms has allowed to accelerate computationally heavy algorithms in multiple fields of science. This paper evaluates the acceleration impact that can be attained for IRLS-based algorithms on a multicore server, showing that an efficient use of the hardware can render a speedup of 10×.

Key words: least squares problems, iterative algorithms, template, multi-core platforms

1 Introduction

Many optimization, parameter estimation, and approximation problems lead to the task of finding an optimal set of parameters $\mathbf{p}_{\text{opt}}$ that minimize the error between the model

$\hat{\mathbf{y}} = f(\mathbf{p})$ and the reality (or observations) contained in $\mathbf{y}$. Mathematically, this can be expressed as

$$\mathbf{p}_{\text{opt}} = \arg\min_{\mathbf{p}} ||f(\mathbf{p}) - \mathbf{y}||. \tag{1}$$

The uniqueness of the solution and the complexity of the parameter estimation depends both on the type of error norm $||.||$ used and the type of function $f(.)$.

Linear least squares problems (LLS) form a special class of problems, where $\hat{\mathbf{y}} = f(\mathbf{p}) = \mathbf{Mp}$ is linear in $\mathbf{p}$, and $\mathbf{M}$ is called the modeling matrix. Moreover, the error to be minimized e_{LS} is the sum of squares of the deviations:

$$e_{\text{LS}} = \sum_{k=1}^{K} |\hat{y}_k - y_k|^2 = ||\mathbf{Mp} - \mathbf{y}||_2 = (\mathbf{Mp} - \mathbf{y})^T(\mathbf{Mp} - \mathbf{y}). \tag{2}$$

We note that for complex valued problems, the Hermitian transpose $(.)^H$ (transpose and complex conjugate) should be used instead of $(.)^T$ in all equations throughout the paper.

Since e_{LS} is a second-order function of $\mathbf{p}$, it has a unique minimum, that is in closed formly:

$$\mathbf{p}_{\text{opt}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{y}. \tag{3}$$

This makes parameter estimation significantly simpler compared to general nonlinear optimization problems. We note that a mathematically equivalent, but numerically better posed method to obtain $\mathbf{p}_{\text{opt}}$ is to solve the set of linear equations given by

$$(\mathbf{M}^T\mathbf{M})\mathbf{p}_{\text{opt}} = \mathbf{M}^T\mathbf{y}. \tag{4}$$

1.1 Iteratively reweighted least squares

Unfortunatelly, not all optimization problems can be expressed as LLS problems, either because the output is not a linear function of the parameters, or the L_2 norm is not appropriate. However, many of these problems can be approximated as a LLS problem with additional weights. In this case, all the deviations $(\hat{y}_k - y_k)$ are multiplied by a constant w_k before the norm is computed. Mathematically, this can be formulated as a multiplication with a diagonal matrix $\mathbf{W} = < \mathbf{w} >$, where $\mathbf{w}$ is the vector containing the elements w_k. In this case, the error is approximated as

$$e = ||f(\mathbf{p}) - \mathbf{y}||^2 \approx e_{\text{WLS}} = ||\mathbf{W}(\mathbf{Mp} - \mathbf{y})||_2^2 = (\mathbf{Mp} - \mathbf{y})^T\mathbf{W}^T\mathbf{W}(\mathbf{Mp} - \mathbf{y}). \tag{5}$$

If the weight matrix $\mathbf{W}$ is known, Eq. (5) can be solved in one step similarly to an unweighted LS problem (see Eq. (3)) by simply multiplying both $\mathbf{M}$ and $\mathbf{p}$ with $\mathbf{W}$:

$$\mathbf{p}_{\text{opt}} = (\mathbf{M}^T\mathbf{W}^T\mathbf{WM})^{-1}\mathbf{M}^T\mathbf{W}^T\mathbf{Wy}. \tag{6}$$

 ISBN: 978-84-608-6082-2

However, when approximating more complex optimization problems, the weights are generally a function of the parameters: $\mathbf{W} = < w(\mathbf{p}) >$. Let us see L_p-norm minimization as an illustrative example [2]. In this case, the error is

$$e_p = \sum_{k=1}^{K} |\hat{y}_k - y_k|^p = e_{\text{WLS}} = \sum_{k=1}^{K} |\hat{y}_k - y_k|^{p-2} |\hat{y}_k - y_k|^2 = \sum_{k=1}^{K} w_k^2 |\hat{y}_k - y_k|^2,$$

where

$$w_k = |\hat{y}_k - y_k|^{\frac{p-2}{2}}. \tag{7}$$

Clearly, w_k depends on the unknown model output $\hat{y}_k$. This dependency can be eliminated via an iterative algorithm where, in the first iteration, $w_k = 1$; and from that on, the previous model output $\hat{y}_k$ is used to update the weights w_k before computing the new optium via Eq. (6). This is the iteratively reweighted least squares (IRLS) algorithm. (Note that for the L_p-minimization case, there exist more sophisticated algorithms with faster convergence [5, 2].)

The IRLS technique has been applied to numerous optimization problems. This includes FIR and IIR filter design in L_p sense [2, 5], the frequency-domain Steiglitz-McBride algorithm [1], magnitude-priority filter design [3], and sparse recovery [4], to name a few.

From a computational point of view, solving Eq. (6) is very expensive from the computational perspective. To evaluate the necessary resources, we have implemented Eq. (6) using routines from LAPACK (Linear Algebra Package) together with BLAS (Basic Linear Algebra Subprograms) [6]. Both are standard software libraries for numerical linear algebra.

2 Implementation with BLAS and LAPACK

The proposed implementation uses the Fortran implementation through a C-interface. Eq. (6) is programmed following these steps:

1. Matrix-Vector Multiplication: $\mathbf{y} = \mathbf{W}\mathbf{y}$, where $\mathbf{W}$ has a size of $m \times m$ and $\mathbf{y}$ has a size of $m \times 1$.
2. Matrix-Matrix Multiplication: $\mathbf{M} = \mathbf{W}\mathbf{M}$, where $\mathbf{M}$ has a size of $m \times n$.
3. Matrix-Matrix Multiplication: $\mathbf{A} = \mathbf{M}^T\mathbf{M}$, where $\mathbf{A}$ has a size of $n \times n$.
4. Matrix-Vector Multiplication: $\mathbf{b} = \mathbf{M}^T\mathbf{y}$, where $\mathbf{b}$ has a size of $n \times 1$.
5. Solving the Linear System: $\mathbf{A}\mathbf{p}_{\text{opt}} = \mathbf{b}$, where $\mathbf{p}_{\text{opt}}$ has a size of $n \times 1$.

Algorithm 1 shows the pseudo-code of the implementation together with the LAPACK and BLAS routines. It is important to point out that matrix $\mathbf{W}$ is a diagonal matrix. Thus,

Algorithm 1 Implementation of an IRLS algorithm

Input: $\mathbf{M}$, $\mathbf{W}$, $\mathbf{y}$, m, n
Output: $\mathbf{p}_{\text{opt}}$

```
1: int info=0;
2: int iptv[m];
3: y=Element-wiseMulti(W,y).
4: M=Element-wiseMulti(W,M).
5: dgemm_(T, N, n, n, m, 1.0, M, n, M, m, 0.0, A, n);
6: dgemv_(T, m, n, 1.0, M, m, y, 1, 0.0, b, 1);
7: dgesv_( n, 1, p_opt, n, iptv, b, n, info);
```

$\mathbf{y} = \mathbf{W}\mathbf{y}$ is carried out by a `for` loop that element-wise multiplies the elements of vector $\mathbf{y}$ with the diagonal values of matrix $\mathbf{W}$, while $\mathbf{M} = \mathbf{W}\mathbf{M}$ is carried out in two `for` loops that element-wise scale the rows of matrix $\mathbf{M}$ with the diagonal values of matrix $\mathbf{W}$. We denote these operations at Algorithm 1 by a function called `Element-wiseMulti`. Note that all LAPACK routines start with `d_`, which indicates that we deal with *double* precision *real* data.

3 Computational Performance of the IRLS

We have tested the proposed implementation on system composed by the following features.

- CPU: Two Intel Xeon CPU E5-2697 at 2.70 GHz.
- CPU cores: 12 cores per CPU
- Hyperthreading: Yes
- Operative System: Linux CentOS release 6.5
- Architecture: x86_64

The implementation have been evaluated for a size of $m = 8000$ and $n = 4000$. Table 1 shows the time employed for each step of the algorithm. Note that steps 3 and 4 from Algorithm 1 are omitted since their computational time is negligible.

As shown in Fig. 1, we achieved a speedup larger than 10 for this problem. Focusing on the computational routines, the operation that requires more time is the `dgemm_`, since it has the highest computational cost in comparison with the other operations.

4 Conclusion and Future work

In this paper, we have experimentally assessed the performance of the IRLS algorithm implemented on a multi-core platform. Our results show that the performance improves

	Number of cores				
	1	4	8	16	24
`dgemm_`	9.846	2.708	1.509	1.088	0.908
`dgemv_`	0.018	0.007	0.006	0.005	0.004
`dgesv_`	1.750	0.485	0.286	0.178	0.160
Total time	11.614	3.200	1.801	1.270	1.073

Table 1: Time in seconds employed for each step of the Algorithm 1 for $m = 8000$ and $n = 4000$

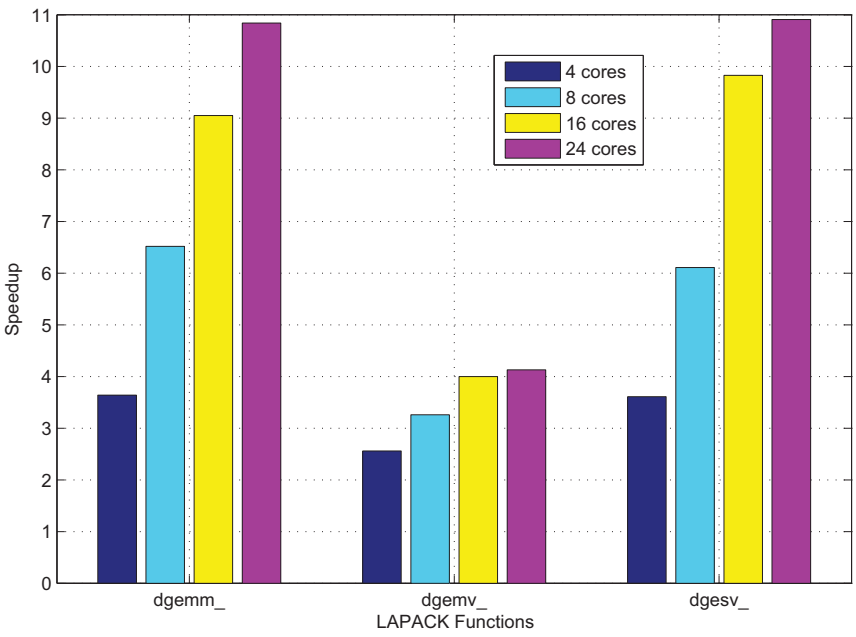


Figure 1: Speedup achieved by each routine that is used in the implementation.

as the number of cores increases. It is important to note that the matrix-matrix operation `dgemm_` consumes most of the time. As a future work, we plan to execute the previous algorithm using *single* precision and complex arithmetic. Moreover, further optimization efforts will be carried out in order to reduce the computational time.

Acknowledgements

This work was conducted in spring 2016 when Jose A. Belloch was a visiting postdoctoral researcher at Budapest University of Technology and Economics thanks to the European Network COST Action IC1305: Network for sustainable Ultrascale computing (NESUS) inside the program Short Term Scientific Mission with the following reference: COST-STSM-ECOST-STSM-IC1305-020416-072431. Different public institutions (Spain and Europe) have also supported this work with projects: TIN2014-53495-R, TEC2013-47141-C4-4-R, TEC2015-67387-C4-1-R, PROMETEOII/2014/003, CAPAP-H5 network and TIN2014-53522-REDT.

References

[1] L. B. Jackson, *Frequency-Domain Steiglitz-McBride Method for Least-Squares Filter Design, ARMA Modeling, and Periodogram Smoothing*, Signal Processing Letters **15** (2008) 49–52.

[2] C. S. Burrus, *Iterative reweighted least squares*, OpenStax-CNC document (2012) http://cnx.org/content/m45285/1.12.

[3] B. Bank, *Magnitude-priority filter design for audio applications*, Proc. 132nd AES Convention (2012) Preprint No. 8591, Budapest, Hungary

[4] I. Daubechies, R. Devire, M. Fornasier, and C. S. Gntrk, *Iteratively reweighted least squares minimization for sparse recovery*, Computer Music J. **23** (2010) 52–69

[5] S. W. Khang, *Best Lp approximation*, Mathematics of Computation **26** (1972) 505–508.

[6] J. J Dongarra, J. D. Croz, S. Hammarling, and R. J. Hanson, *A proposal for an extended set of Fortran basic linear algebra subprograms*, ACM Signum Newsletter **20** (1985) 2–18.

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

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Conversational recommendation to avoid the cold-start problem

F. Benito-Picazo[1], M. Enciso[1], C. Rossi[1] and A. Guevara[1]

[1] *Department of Languages and Computer Science, Universidad de Málaga, Andalucía Tech, Málaga, Spain*

emails: fbenito@lcc.uma.es, enciso@lcc.uma.es, rossi@uma.es, guevara@uma.es

Abstract

Recommender systems has become a widespread topic, allowing to connect user demands to those products more suitable to their preferences. The more information we provide to the system, the better the system works. This is a weak point of recommenders: they need an initial information belonging to each new user. In this paper we propose to avoid the so-called cold-start problem by using a conversational recommendation approach. We consider products characteristics as attributes and deal with the attribute implications by means of the simplification logic to guide the user in the search.

Key words: Recommendation systems, conversational recommendation, logic, implications

1 Introduction

Nowadays recommender systems have established a solid field of knowledge within information technologies. They are a kind of software that group together a wide range of techniques and applications with the aim of providing the best user experience [18]. There has been much progress done towards recommender systems during last decade [1] but there is still so much work remained. Examples of the applications concerning recommender systems go over many different topics of today's society such as recommending books, music, documents, e-commerce, tourism, medical diagnosis, among others. Recommender Systems constitute a hot topic indeed, as we can notice by the way in which many top companies worldwide spend their efforts and resources developing more and better systems for them to

increase their benefits. By these means, companies with absolutely different market niches delegate their most important duties to recommender systems due to the wide range of possibilities they offer; and yet companies selling products are not all of them, global leaders in other fields as totally come aboard with recommender systems by recommending new friends, groups, followers, and other social connections.

When recommendations are based on of the element evaluation made by other users or by similarity between the user preferences and these characteristics, recommender systems need to face many problems before they can flow into good recommendations. The first one we need to remark is the well-known cold-start problem [10], that appears when recommender systems try to elaborate reliable recommendations from the absence of initial information. Cold-start problem may be handled by requesting other agents to share what they have already learned from their respective users [11]. Also, new items (those which have not received any ratings from the community yet) would be assigned a rating automatically, based on those given by the community to other similar items [20] and so, we are at the mercy of similarity rules. In the same direction, until the new element has not been evaluated by a significant number of users, the system will not be able to recommend it. An item that is not recommended remain unnoticed by most of the user community, thus, we can enter into a vicious circle in which a set of elements of the recommender systems will be left out of the rating process and/or recommendations continuously [16]. In most of cases, users do not rate all the features we would desire for the optimum running of the recommender systems, this reveals scarcity problem.

In this work, we propose to deal with the cold-star problem by introducing an information flow based on the dialogue with the user. The lack of initial information is avoided with the design of a process with allows to collect this information of the user and storing them for further access to the system. This process, as we shall see, is a recommender-like system, to allow the user for getting some usefullness in its use.

2 Recommender systems and the conversational issue

There exists different kinds of recommender systems usually classified on how recommendations are made [1]. The most known and extended ones are collaborative filtering, content-based and demographic systems. Besides, in recent years there has been a great expansion of context-aware recommender systems [2] and knowledge-based recommender systems [14]. Other group of recommender systems that worths to be considered is that one focused on recommendations involving group of people [9]. Collaborative filtering systems [13], recommend items that other users have already rated before. Recommendations made by content-based systems present items similar to the ones the user preferred in the past [12]. Context-aware recommender systems try to adapt their recommendations to the world around the user. Finally, knowledge-based approaches are different; they manage functional

knowledge about how an item matches a particular need, and they can therefore reason about the relationship between a need and a possible recommendation. These characteristics make knowledge-base recommender systems not only valuable systems on their own, but also highly complementary to other types of recommender systems. However, the history of recommender systems has broadly demonstrated that best strategies are those who merge characteristics from different kinds of recommender systems in order to generate hybrids conforming best features of each one [6, 4].

In general, most of widely used recommendation techniques requires information to build a user profile before generating a result. In some cases, that information may be gathered explicitly: for example, requiring data about age, gender, etc. during a registration process, or by means of ratings and opinions about the recommended items. In other cases, the system may get implicit information from the browsing and/or purchase user history.

Nevertheless, there are contexts in which this previous information it is not available. This is the case of the well-known cold-start problem, when a new user asks for his first recommendation and obviously the system has not any information about him. This situation also occurs in systems where users make occasional use.

An interesting approach to solve this problem us the use of the so-called conversational recommender systems [7, 8]. These are closely related with critiquing recommender systems [17, 21]. In these works, recommendation is enriched by means of a dialog with the user that allows an incremental elicitation of his preferred item features. To promote an effective use of this approach, our proposal produces as an output a recommendation only based on the user dialogue information. In this way, the system is attractive for those user that are new in the system and can be used as a preliminary system to store user preferences for further accesses.

3 A logic approach to conversational recommendation

Our proposal to integrate recommender systems and the conversational issue is based on a sound and complete logic. As we shall see, such an strong basis allows us to include a reasoning method in the process and allows us to store the information in a natural way to be managed in the future by knowledge-base recommenders.

We built our framework on a basic elements, the implications. They correspond to formulas $a_1 \wedge \ldots \wedge a_n \rightarrow b_1 \wedge \ldots \wedge b_m$. The propositions $a_1, \ldots a_n, b_1, \ldots, b_m$ are elements of a set Ω and they are interpreted as properties concerning attributes. For this reason, propositional symbols are named attributes. To compact notation it is usual to denote the above formulas as $A \rightarrow B$ being $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_m\}$ i.e. sets of attributes are conjunctively interpreted.

The symbolic management of implications was originally proposed in [3]. However, due to the central role that transitivity plays in this axiomatic system, the development

of executable method to solve implications problems has rest on indirect methods. For instance, the proposal to solve the attribute closure, i.e. to find the maximal set of attributes A^+ such that the implication $A \rightarrow A^+$ holds has been traditionally tackle by using a basic method which exhaustively uses the subset relation to add new elements in the conclusion.

The introduction of the Simplification Logic, named $\mathbf{SL}_{FD}$, [5] opened the door to the development of automated reasoning methods directly based on its novel axiomatic system. $\mathbf{SL}_{FD}$ considers reflexivity as axiom scheme

$$\texttt{[Ref]} \quad \frac{}{A \rightarrow A}$$

together with the following inference rules called Fragmentation, Composition and Simplification respectively.

$$\texttt{[Frag]} \ \frac{A \rightarrow BC}{A \rightarrow B} \qquad \texttt{[Comp]} \ \frac{A \rightarrow B,\ C \rightarrow D}{AC \rightarrow BD} \qquad \texttt{[Simp]} \ \frac{A \rightarrow B,\ C \rightarrow D}{A(C \setminus B) \rightarrow D}$$

Later, in [15] we presented an attribute closure method closely tied to the Simplification logic axiomatic system. Apart from having a strong base, the main advantage of our method is that its output is twofold: besides the maximal set constituting the closure of the input, it also renders a reduced set of implications which enclose the semantics that is outside the set A^+. We would like to remark that this two inputs are computed in linear time, overtaking the hard cost of a data mining process if we were interested in extracting the new set of implication for the reduced dataset after each search step.

This characteristics provides a key information to further inferences in an iterative search process. This is the core of our proposal to design a conversational recommendation based on our attribute closure operator. The recommendation process will go along the following points:

0. We depart from the premise that we have a dataset containing items and attributes, and the set of implications that holds on it. This is considered point zero and, as we have mentioned, it does not requieres any information from the user to be started.

1. Once we count on this information, the user interacts with the system by selecting an attribute we wish an item to fit.

2. Then, the process flows into the closure algorithm calculating both the set closure for this attribute and above all, the set of implications that remains outside the closure and complete them.

3. Once the closure algorithm has finished, a new reduced dataset is shown. At this point, we can stop the interaction whether we are already satisfied with the result or we can go ahead trying to get a more suitable recommendation. The improvement here goes as follows. For further queries, we have reduced the number of available attributes

deleting those included in the closure set. Even that this could be accomplished by classic closure algorithms, the major point of our method is that, *at the same time*, we also reduce the number of implications, and so, in every refining-attempt we do not need to start the process form the beginning but continuing from here, where both attributes and implications have been decreased. Consequently, the process maintains its linear complexity and the interaction becomes truly faster.

4. In this way, we select a new attribute and resume the search.

5. We carry on selecting attributes until we get a satisfying recommendation or we run out of attributes.

4 Conclusion and future works

In this paper we propose to approach the cold-start problem. We mining the dataset containing the product information to get a set of attribute implications. This set is managed by using the inference system of simplification logic to guide the search of new users.

As a future work, we propose to study the impact of simplification closure in the performance of our approach. Our method allows to get, in an iterative way, intermediate closure set of attributes and the corresponding reduced set of implications. This characteristics allows to proceed step by step and, at the same time, accelerate the search.

Acknowledgements

This work is partially supported by project TIN2014-59471-P of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, June 2005.

[2] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In Ricci et al. [19], pages 217–253.

[3] W. W. Armstrong. Dependency structures of data base relationships. In *IFIP Congress*, pages 580–583, 1974.

[4] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutirrez. Recommender systems survey. *Knowledge-Based Systems*, 46(0):109 – 132, 2013.

[5] P. Cordero, M. Enciso, A. Mora, and I. P. de Guzmán. $\mathrm{Sl}_{\mathrm{fd}}$ logic: Elimination of data redundancy in knowledge representation. *IBERAMIA*, pages 141–150, 2002.

[6] L. M. de Campos, J. M. Fernndez-Luna, J. F. Huete, and M. A. Rueda-Morales. Combining content-based and collaborative recommendations: A hybrid approach based on bayesian networks. *International Journal of Approximate Reasoning*, 51(7):785 – 799, 2010.

[7] S. Guerrero and M. Salamo. Increasing retrieval quality in conversational recommenders. *IEEE Transactions on Knowledge and Data Engineering*, 24(10):1876–1888, 2012.

[8] D. Jannach and G. Kreutler. Rapid development of knowledge-based conversational recommender applications with advisor suite. *Journal of Web Engineering*, 6(2):165–192, 2007.

[9] H.-N. Kim and A. El-Saddik. A stochastic approach to group recommendations in social media systems. *Information Systems*, 50(0):76 – 93, 2015.

[10] H.-N. Kim, A. El-Saddik, and G. Jo. Collaborative error-reflected models for cold-start recommender systems. *Decision Support Systems*, 51(3):519–531, 2011.

[11] Y. Lashkari, M. Metral, and P. Maes. Collaborative interface agents. In *In Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 444–449. AAAI Press, 1994.

[12] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In Ricci et al. [19], pages 73–105.

[13] M. Maleszka, B. Mianowska, and N. T. Nguyen. A method for collaborative recommendation using knowledge integration tools and hierarchical structure of user profiles. *Knowl.-Based Syst.*, 47:1–13, 2013.

[14] M. Mandl, A. Felfernig, E. Teppan, and M. Schubert. Consumer decision making in knowledge-based recommendation. *Journal of Intelligent Information Systems*, 37:1–22, 2011.

[15] A. Mora, P. Cordero, M. Enciso, I. Fortes, and G. Aguilera. Closure via functional dependence simplification. *Int. J. Comput. Math.*, 89(4):510–526, 2012.

[16] S.-T. Park and W. Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 21–28, New York, NY, USA, 2009. ACM.

[17] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth. Incremental critiquing. *Knowledge.-Based Systems*, 18(4-5):143–151, 2005.

[18] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[19] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[20] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[21] W. Trabelsi, N. Wilson, D. Bridge, and F. Ricci. Preference dominance reasoning for conversational recommender systems: a comparison between a comparative preferences and a sum of weights approach. *International Journal on Artificial Intelligence Tools*, 20(4):591–616, 2011.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Numerical solution of two–dimensional nonlinear Volterra integral equations

M. I. Berenguer[1] and D. Gámez[1]

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

emails: maribel@ugr.es, domingo@ugr.es

Abstract

Using fixed-point techniques and Faber–Schauder systems in adequate Banach spaces, allows us to design a numerical algorithm in order to solve and important type of two-dimensional nonlinear integral equation.

Key words: Two dimensional integral equations, Schauder bases, Banach spaces, Fixed Point Theorem, numerical methods.
MSC 2000: AMS 45A05, 45L05, 45N05,65R20.

1 Introduction

Several problems in mathematics, physics, engineering and other disciplines can be modeled using two-dimensional integral equations. These are usually difficult to solve analytically and in many cases the solution must be approximated. Therefore, in recent years several numerical approaches have been proposed. For example, a numerical method based on Haar wavelet (see [1]), two-dimensional orthogonal triangular functions and the two-dimensional rationalized Haar are used to approximate the solution (see [2] and [3], respectively), a differential transformation method (see [9]), a numerical scheme based on the moving least squares method (see [11]), a operational matrix and two-dimensional block pulse functions methods (see [12]), a method using Legendre polynomials (see [13]), a method based in the developing the two–dimensional differential transform for double integrals (see [16]).

 ISBN: 978-84-608-6082-2

2 The problem

The aim of this work is to introduce a new numerical method in order to approximate the solution of the two–dimensional nonlinear integral equations of Volterra:

$$f(s,t) = g(s,t) + \int_{\gamma}^{t} \int_{\alpha}^{s} K(s,t,x,y,f(x,y))\, dxdy \tag{1}$$

where $\Omega = [\alpha, \alpha+\beta] \times [\gamma, \gamma+\delta]$, $f \in C(\Omega)$ is the solution to be approximated, and g and K are given real-valued continuous funtions defined, respectively, on Ω and $W = \{(s,t,x,y,z) \in \mathbb{R}^5 : \alpha \leq x \leq s < \alpha+\beta,\ \gamma \leq y \leq t < \gamma+\delta\}$.

Using fixed-point techniques and Faber–Schauder systems in adequate Banach spaces, we present an efficient analytical and numerical procedure for solving two–dimensional nonlinear Volterra integral equations. Such tools have been used successfully in the study of certain types of one–dimensional integral, integro-differential and differential equations (see [4], [5], [6]). The study of the convergence and error will be done. The efficiency of this method will be shown with numerical results.

Acknowledgements

This work has been partially supported by Junta de Andalucía Grant FQM359 and the E.T.S.I.E. of the University of Granada (Spain).

References

[1] I. AZIZ, S. ISLAM AND F. KHAN, *A new method based on Haar wavelet for the numerical solution of two-dimensional nonlinear integral equations*, J. Comput Appl. Math. **272** (2014) 70–80.

[2] E. BABOLIAN, K. MALEKNEJAD, M. ROODAKI AND H. ALMASIEH, *Two-dimensional triangular functions and their applications to nonlinear 2D Volterra-Fredholm integral equations*, Comput. Math. Appl. **60** (2010) 1711–1722.

[3] E. BABOLIAN, S. BAZM AND P. LIMA, *Numerical solution of nonlinear two-dimensional integral equations using rationalized Haar functions*, Commun Nonlinear Sci. **16** (2011) 1164–1175.

[4] M. I. BERENGUER, D. GAMEZ, A. I. GARRALDA GUILLEM AND M. C. SERRANO PEREZ, *Nonlinear Volterra integral equation of the second kind and biorthogonal systems*, Abstr. Appl. Anal. Vol. **2010** (2010) 11 pages.

[5] M. I. BERENGUER, D. GAMEZ AND A. J. LOPEZ LINARES, *Fixed–point iterative algorithm for the linear Fredholm–Volterra integro–differential equations*, J. Appl. Math. Vol. **2012** (2012) 12 pages.

[6] E. CASTRO, D. GAMEZ, A. I. GARRALDA GUILLEM AND M. RUIZ GALAN, *High order linear initial–value problems and Schauder bases*, Appl. Math. Model. **31** (2007) 2629–2638.

[7] B. R. GELBAUM AND J. GIL DE LAMADRID, *Bases of tensor products of Banach spaces*, Pacific. J. Math. **11** (1961) 1281–1286

[8] M. HADIZADEH AND N. MOATAMEDI, *A new differential transformation approach for two-dimensional Volterra integral equations*, Int. J. Comput. Math. Vol. **84** (2007) 515–526.

[9] B. JANG, *Comments on "Solving a class of two-dimensional linear and nonlinear Volterra integral equations by the differential transform method"*, J. Comput Appl. Math. **233** (2009) 224–230.

[10] G. J. O. JAMESON, *Topology and Normed Spaces*, Chapman-Hall, London, 1974.

[11] D. MIRZAEI AND M. DEHGHAN, *A meshless based method for solution of integral equations*, Appl. Numer. Math. **60** (2010) 245–262.

[12] S. NAFAJALIZADEH AND R. EZZATI, *Numerical methods for solving two-dimensional nonlinear integral equations of fractional order by using two-dimensional block pulse operational matrix*, Appl. Math. Comput. **280** (2016) 46–56.

[13] S. NEMATI, P. M. LIMA AND Y. ORDOKHANI, *Numerical solution of a class of two-dimensional nonlinear Volterra integral equations using Legendre polynomials*, J. Comput Appl. Math. **242** (2013) 53–69.

[14] Z. SEMADENI, *Schauder Bases in Banach Spaces of Continuous Functions*, Springer–Verlag, Berln, 1982.

[15] Z. SEMADENI, *Product Schauder bases and approximation with nodes in spaces of continuous functions*, Bull. Acad. Polon. Sci. **11** (1963) 387–391.

[16] A. TARI, M. Y. RAHIMI, S. SHAHMORAD AND F. TALATI, *Solving a class of two–dimensional linear and nonlinear Volterra integral equations by the differential transform method*, J. Comput Appl. Math. **228** (2009) 70–76.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An efficient method for solving two-dimensional Fredholm integral equations

M. I. Berenguer[1] and D. Gámez[1]

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

emails: maribel@ugr.es, domingo@ugr.es

Abstract

In this work we obtain an approximation of the solution of the two-dimensional Fredholm integral equations using two analytical tools: the Banach fixed point theorem and the propierties of a biorthogonal system in a Banach space.

Key words: Two dimensional integral equations, Schauder bases, Banach spaces, Fixed Point Theorem, numerical methods.
MSC 2000: AMS 45A05, 45L05, 45N05,65R20.

1 Introduction

The two dimensional integral equations are widely used for solving many problems in mathematics, physics and engineering. On many occasions it is not possible to find an exact solution to these equations. Therefore, in recent years several numerical approaches have been proposed. For example, a method based on interpolation by Gaussian radial basis function (see [1]), a numerical method based on Haar wavelet (see [2]), a numerical method using rationalized Haar functions (see [3]), a spectral meshless radial point interpolation method (see [7]), a discrete Galerkin and iterated Galerkin method (see [9]), a integral mean value method (see [10]), a piecewise interpolating polynomial technique (see [12]), a two-dimensional modification of hat functions method (see [13]), an iterative numerical method of succsesive approximations (see [14]).

2 The problem

The purpose of this work is to develop an effective method for approximating the solution of the two–dimensional nonlinear Fredholm integral equations:

$$f(s,t) = g(s,t) + \int_{\gamma}^{\gamma+\delta} \int_{\alpha}^{\alpha+\beta} K(s,t,x,y,f(x,y))\, dxdy \tag{1}$$

where $\Omega = [\alpha, \alpha+\beta] \times [\gamma, \gamma+\delta]$, $f \in C(\Omega)$ is the solution to be approximated, and g and K are given real-valued continuous funtions defined, respectively, on Ω and $\Omega \times \Omega \times \mathbb{R}$.

The numerical method to solve the above mentioned equation is based in two analytical techniques: the Banach fixed point theorem and Schauder bases in Banach space. Such tools have been used successfully in the study of certain types of one–dimensional integral, integro-differential and differential equations (see [4], [5], [6]).

We recall some auxiliary facts from the Banach fixed point theorem and the Theory of Schauder Bases, we present the method, getting and explicit control of the error committed. Finally we illustrate the theoretical results with some examples.

Acknowledgements

This work has been partially supported by Junta de Andalucía Grant FQM359 and the E.T.S.I.E. of the University of Granada (Spain).

References

[1] A. ALIPANAH AND S. ESMAELI, *Numerical solution of the two-dimensional Fredholm integral equations using Gaussian radial basis function*, J. Comput Appl. Math. **235** (2011) 5342–5347.

[2] I. AZIZ, S. ISLAM AND F. KHAN, *A new method based on Haar wavelet for the numerical solution of two-dimensional nonlinear integral equations*, J. Comput Appl. Math. **272** (2014) 70–80.

[3] E. BABOLIAN, S. BAZM AND P. LIMA, *Numerical solution of nonlinear two-dimensional integral equations using rationalized Haar functions*, Commun Nonlinear Sci. **16** (2011) 1164–1175.

[4] M. I. BERENGUER, M. V. FERNÁNDEZ MUNOZ, A. I. GARRALDA GUILLEM AND M. RUIZ GALÁN, *Numerical treatment of fixed point applied*

to nonlinear Fredholm integral equation, Fixed Point Theory A. Vol. **2009** (2009) 8 pages.

[5] M. I. BERENGUER, M. V. FERNÁNDEZ MUOZ, A. I. GARRALDA GUILLEM AND M. RUIZ GALÁN, *A sequential approach for solving the Fredholm integro-differential equation*, Appl. Numer. Math. **62** (2012) 297–304.

[6] E. CASTRO, D. GAMEZ, A. I. GARRALDA GUILLEM AND M. RUIZ GALAN, *High order linear initial–value problems and Schauder bases*, Appl. Math. Model. **31** (2007) 2629–2638.

[7] H. FATAHI, J. SABERI-NADJAFI AND E. SHIVANIAN, *A new spectral meshless radial point interpolation (SMRPI) method for the two-dimensional fredholm integral equations on general domains with error analysis*, J. Comput Appl. Math. **294** (2016) 196–209.

[8] B. R. GELBAUM AND J. GIL DE LAMADRID, *Bases of tensor products of Banach spaces*, Pacific. J. Math. **11** (1961) 1281–1286

[9] H. GUOQIANG AND R. Wang, *Richardson extrapolation of iterated discrete Galerkin solution for two-dimensional Frredholm integral equations*, J. Comput Appl. Math. **139** (2002) 49–63.

[10] M. HEIDARI, Z. AVAZZADEH, H. NAVABPOUR AND G. B. LOGHMANI, *Numerical solution of Fredholm integral equations of the second kind by using integral mean value theorem II. High dimensional problems*, Appl. Math. Model. **37** (2013) 432–442.

[11] G. J. O. JAMESON, *Topology and Normed Spaces*, Chapman-Hall, London, 1974.

[12] F. LIANG AND F.-R. LIN, *A fast numerical solution method for two dimensional Fredholm integral equations of the second kind based on piecewise polynomial interpolation*, Appl. Math. Comput **233** (2010) 3073–3088.

[13] F. MIRZAEE AND E. HADADIYAN, *Numerical solution of linear integral equations via two-dimensional modification of hat functions* , Appl. Math. Comput. **250** (2015) 805–816.

[14] S. M. SADATRASOUL AND R. EZZATI, *Numerical solution of two-dimensional nonlinear Hammerstein fuzzy integral equations based on optimal fuzzy quadrature formula*, J. Comput Appl. Math. **292** (2016) 430–446.

[15] Z. SEMADENI, *Schauder Bases in Banach Spaces of Continuous Functions*, Springer–Verlag, Berln, 1982.

[16] Z. SEMADENI, *Product Schauder bases and approximation with nodes in spaces of continuous functions*, Bull. Acad. Polon. Sci. **11** (1963) 387–391.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Advances in time-dependent current-density functional theory

Arjan Berger[1]

[1] *Laboratoire de Chimie et Physique Quantiques, IRSAMC, Université Toulouse III - Paul Sabatier, CNRS and European Theoretical Spectroscopy Facility (ETSF), 118 Route de Narbonne, 31062 Toulouse Cedex, France, University of First and Second Authors*

emails: arjan.berger@irsamc.ups-tlse.fr

Abstract

In this talk I will discuss a solution to the problem of the gauge dependence of molecular magnetic properties (magnetizabilities, circular dichroism) using time-dependent current-density functional theory [1]. I will also present a new parameter-free functional that accurately describes the optical absorption spectra of insulators, semiconductors and metals [2].

References

[1] N. Raimbault, P.L. de Boeij, P. Romaniello, and J.A. Berger, Phys. Rev. Lett. **114**, 066404 (2015) .

[2] J.A. Berger, Phys. Rev. Lett. **115**, 137402 (2015) .

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Parallelization of the 3D Fast Wavelet Transform on a Cluster of Raspberry Pi 2 Boards

Gregorio Bernabé[1], Raúl Hernández[1] and Manuel E. Acacio[1]

[1] *Computer Engineering, University of Murcia*

emails: gbernabe@ditec.um.es, raulhnpacheco@hotmail.com, meacacio@ditec.um.es

Abstract

Current low-cost general-purpose single-board computing (SDC) devices are gaining increasing interests in research computing due to their very low cost/performance ratio. Among all the SDCs available nowadays, Raspberry Pi devices constitute maybe the most renowned representative. On the other hand, the wavelet transform plays an important role in contemporary standards for image compression (such as JPEG-2000) and video compression (MPEG-4). In this work we evaluate two parallelization strategies of the 3D fast wavelet transform (3D-FWT), both implemented using MPI, on a cluster of Raspberry Pi 2 SDCs. We find out that while noticeable speed-ups can be obtained when all MPI processes run on cores of a single Raspberry Pi 2 SDC, performance drops drastically when they spread to several boards. The reason for this is the limited bandwidth that the on-board LAN port can deliver, and that proves insufficient for the fine-grained, high-volume communication requirements of the two parallelization strategies.

Key words: General-purpose single-board computing (SDC); Raspberry Pi 2; 3D fast wavelet transform (3D-FWT); Parallelization strategies; Speed-up

1 Introduction

Continuous improvements in the technologies used to build computers have recently made possible the fabrication of extremely low-cost general-purpose single-board computing devices. Nowadays, one can buy one of these *tiny* computers for a few dollars and make it run Windows 10 or Ubuntu-Linux operating systems [21]. Today, one of the most renowned examples of these single-board computers is Raspberry Pi devices. Although the initial aim of these devices was to promote the teaching of basic computer science in schools [14] and

developing countries [15], recent appearance of single-board computers (SBC) with multi-core CPU chips have attracted interest of a multitude of projects trying to take advantage of their very low cost-performance ratio (i.e. in data centers [23]).

Raspberry Pi SBCs are priced between US$20 and US$35[1]. The development of the Raspberry Pi SDCs has gone through several generations. The first generation (Raspberry Pi 1) was released in February 2012 in basic model A and a higher specification model B. A+ and B+ models were released a year later. All these Raspberry Pi's had a single-core ARM-based CPU chip which restricted their use to *toy* computers. Raspberry Pi 2 model B was released in February 2015 and its main novelty was the introduction of a 4-core ARM-based CPU chip, which brought a 6-fold performance increase compared to their predecessor. The last family member, the Raspberry Pi 3 model B, was launched in February 2016 and adds wireless connectivity (2.4 GHz WiFi 802.11n and Bluetooth 4.1).

In the last few years, a very attractive area of research involves the proposal and evaluation of different transform functions that may overcome the limitations that the discrete cosine transform (DCT) used by MPEG-2 presents for some particular types of video. Wavelet techniques have recently generated much interest in applied areas and the wavelet transform has been mainly applied to images. Several coders have been developed using the 2D wavelet transform [1, 17, 25]. The latest image compression standard, JPEG-2000 [19, 22], is also based on the 2D wavelet transform with a dyadic mother wavelet transform. The 3D wavelet transform has been also applied for compressing video. Since one of the three spatial dimensions can be considered similar to time, Chen and Pearlman developed a three-dimensional subband coding to code video sequences [8], later improved with an embedded wavelet video coder using 3D set partitioning in hierarchical trees (SPHIT) [16]. Today, the standard MPEG-4 [2, 3] supports an ad-hoc tool for encoding textures and still images, based on a wavelet algorithm. In a previous work [5], we have presented the implementation of a lossy encoder for medical video based on the 3D fast wavelet transform. This encoder achieves both high compression ratios and excellent quality, so that medical doctors cannot find longer differences between the original and the reconstructed video. Furthermore, the execution time achieved by this encoder allows for real-time video compression and transmission. In the case of parallelizing the 2D or 3D fast wavelet transform, we have shown previously that contemporary GPGPUs can achieve dramatical speed-ups [7, 10, 11, 12]. Unfortunately, automatic parallelization methods do not yield enough benefits, while manual parallelization methods pose a considerable burden in software development [4].

In this work we study the parallelization of the 3D fast wavelet transform (3D-FWT) on a cluster of Raspberry Pi 2 SDCs. Particularly, we have implemented two different parallelization strategies for the 3D-FWT (decomposition by rows and decomposition by rows and columns) using MPI [13], and we have evaluated them on a cluster of 4 Raspberry

[1] A Pi Zero with smaller footprint and limited IO (GPIO) capabilities was released in November 2015 for US$5.

Pi 2 SDCs. Our results show that while noticeable speed-ups can be obtained when all MPI processes run on cores of a single Raspberry Pi 2 SDC, performance drops drastically when they spread to several boards. The reason for this is the limited bandwidth that the on-board LAN port (rated as 10/100 Fast Ethernet, and driven through the on-board USB 2.0 bus) can deliver, and that proves insufficient for the fine-grained, high-volume communication requirements of the two parallelization strategies.

The rest of the paper is organized as follows. Section 2 contains important background about the wavelet transform. The parallelization strategies that we have implemented and evaluated in this work are explained in Section 3. In Section 4 we give the details of the cluster of Raspeberry Pi 2 SDCs used for the evaluation, and then, we present the results. Finally, Section 5 concludes the paper and draws some lines of future work.

2 Wavelet Transform Foundations

The basic idea of the wavelet transform is to represent any arbitrary function f as a weighted sum of functions, referred to as wavelets. Each wavelet is obtained from a mother wavelet function by conveniently scaling and translating it. The result is equivalent to decomposing f into different scale levels (or layers), where each level is then further decomposed with a resolution adapted to that level.

In a multiresolution analysis, there are two functions: the mother wavelet and its associated scaling function. Therefore, the wavelet transform can be implemented by quadrature mirror filters (QMF), $G = g(n)$ and $H = h(n)$, $n \epsilon Z$. H corresponds to a low-pass filter, and G is a high-pass filter. For a more detailed analysis of the relationship between wavelets and QMF see [18].

The filters H and G correspond to one step in the wavelet decomposition. Given a discrete signal, s, with a length of 2^n, at each stage of the wavelet transformation the G and H filters are applied to the signal, and the filter output downsampled by two, thus generating two bands, G and H. The process is then repeated on the H band to generate the next level of decomposition, and so on. This procedure is referred to as the 1D fast wavelet transform (1D-FWT).

It is not difficult to generalize the one-dimensional wavelet transform to the multi-dimensional case [18]. The wavelet representation of an image, $f(x, y)$, can be obtained with a pyramid algorithm. It can be achieved by first applying the 1D-FWT to each row of the image and then to each column, that is, the G and H filters are applied to the image in both the horizontal and vertical directions. The process is repeated several times, as in the one-dimensional case. This procedure is referred to as the 2D fast wavelet transform (2D-FWT).

As in 2D, we can generalize the one-dimensional wavelet transform for the three-dimensional case. Instead of one image, there is now a sequence of images. Thus, a new

dimension has emerged, the time (t). The 3D-FWT can be computed by successively applying the 1D wavelet transform to the value of the pixels in each dimension.

Based on previous work [6], we consider Daubechie's W_4 mother wavelet [9] as an appropriate baseline function. This selection determines the access pattern to memory for the entire 3D-FWT process. Let us assume an input video sequence consisting of a number of frames (3^{rd} dimension), each composed of a certain number of rows and columns (1^{st} and 2^{nd} dimension). The 1D-FWT is performed across all frames for each row and column, that is, we apply the 1D-FWT $rows \times cols$ times in the third dimension. The first 1D-FWT instance requires four elements to calculate the first output element for the reference video and the detailed video, with these elements being the first pixel belonging to the first four frames. The second output element for the reference and detailed video are calculated using the first pixel of the third, fourth, fifth and sixth video frames. We continue this way until the entire reference and detailed video are calculated, and these data are the input used for the next stage.

The 2D-FWT is performed $frames$ times, i.e., once per frame. This transform is performed by first applying the 1D-FWT on each row (*horizontal filtering*) of the image, followed by the 1D-FWT on each column (*vertical filtering*). The fact that *vertical filtering* computes each column entirely before advancing to the next column, forces the cache lines belonging to the first rows to be replaced before the algorithm moves on to the next column. Meerwald et al. [20] propose two techniques to overcome this problem: row extension and aggregation or tiling.

Other studies [24, 26] also report remarkable improvements when applying the *tiling* technique over the 2D-FWT algorithm. Our experience implementing on a CPU the sequential 2D-FWT algorithm revealed a reduction of almost an order of magnitude in the overall execution time with respect to a baseline version. This process can straightforwardly be applied to the 3D case, where reports solid gains on execution times as well, which ranges from 3x to 7x factors [5].

In previous works [10, 11], we contributed with a CUDA implementation for the 2D-FWT running more than 20 times faster than a sequential C version on a CPU, and more than twice faster than optimized OpenMP and pthreads versions implemented on a quad-core CPU. We extended the analysis to the 3D-FWT scenario [7, 12], where speed-up factors have been improved using a new set of optimization techniques. We presented different alternatives and programming techniques for an efficient parallelization of the 3D fast wavelet transform on multicore CPUs and manycore GPUs. OpenMP and pthreads were used on the CPU to build different implementations in order to maximize parallelism, whereas CUDA and OpenCL were selected for data parallelism exploitation on the GPU with an explicit memory handling. Speed-ups of the CUDA version on Fermi architecture were the highest obtained, improving the execution times on CPU on a range from 5.3x to 7.4x for different image sizes, and up to 81 times faster when communications are neglected. Meanwhile,

OpenCL obtains solid gains in the range from 2x for small frame sizes to 3x for larger ones.

3 Parallelization on a cluster of Raspberry Pis

In this Section we present the two parallelization strategies we have considered in this work. In both cases, we have implemented them using MPI.

3.1 Decomposition by rows

Our first parallelization strategy tries to take advantage of the fact that computations on the different rows of each frame in a video sequence can be done in parallel. To do so, in this approach different blocks of rows of each frame of the video sequence are divided among the different participating MPI processes. Then, the 1D-FWT in the time and x dimensions are executed by each process over its assigned rows. Next, every process sends its results to the first process (in our case, process with id 0), which is in charge of applying the 1D-FWT to all the columns of the frame. The low-pass outputs of the 3D-FWT are sent from the first process to the rest of them in order to apply the second iteration of the 3D-FWT and so on. Figure 1 describes graphically how this first parallelization strategy would proceed.

3.2 Decomposition by rows and columns

In an attempt to try to increase the amount of parallel work, we also consider the decomposition by rows and columns parallelization strategy. This is an evolution of the previous decomposition strategy. Particularly, once the 1D-FWT is applied in the time and x dimensions by each process, instead of sending the results to a particular process, in this alternative results are exchanged between all participating processes. Once this done, the 1D-FWT is applied by each process for a block of columns. Then, the low-pass output of the 3D-FWT are also exchanged between all the processes to apply the second iteration of the 3D-FWT and so on. Figure 2 describes graphically how this second parallelization strategy would proceed. It is clear that in this case more work is done in parallel but at the expense of increasing communication requirements.

4 Experiments

We have built a cluster which is composed by four Raspberry Pi 2 Model B nodes, as we can observe in Figure 3. Each node contains a 900 MHz quad-core ARM Cortex-A7 CPU and 1 Gbyte of RAM memory. The interconnection network of the cluster is Ethernet at 100 Mbps. The Operating System is Raspbian Wheezy. In our cluster we have installed MPICH2 (v1.4.1) as the MPI library implementation.

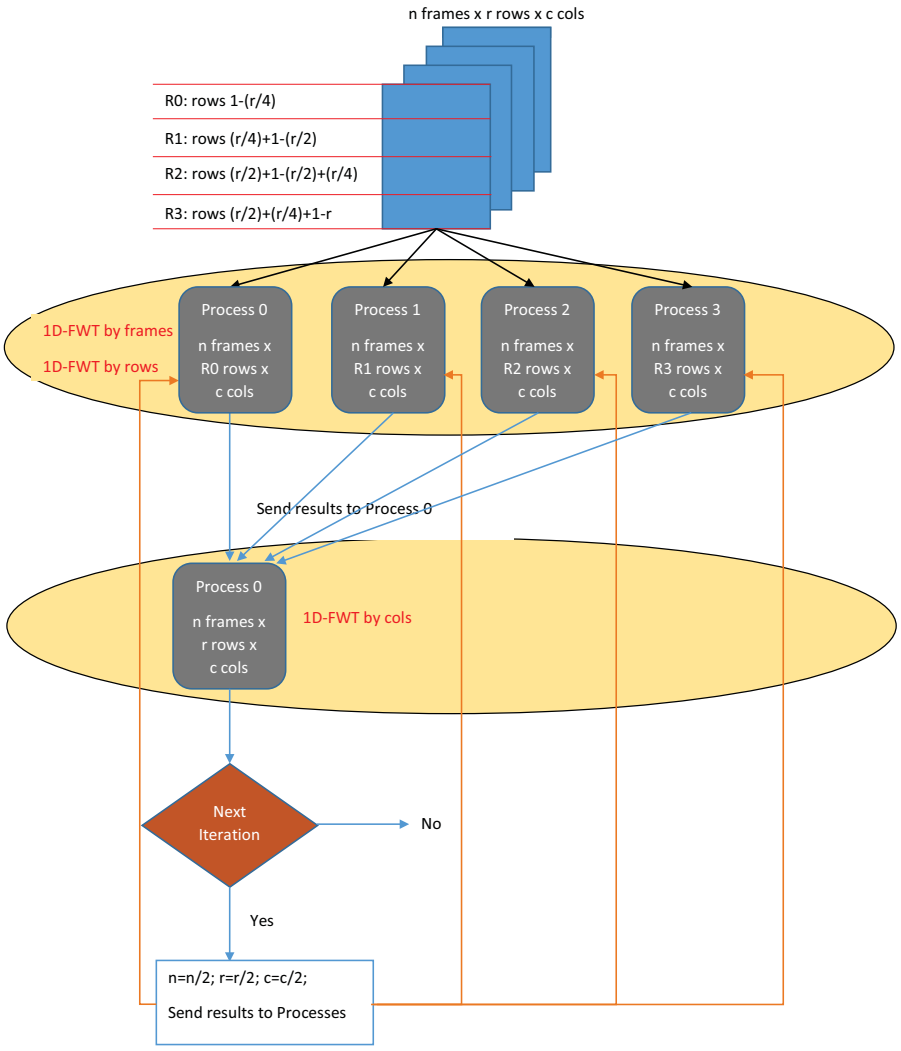


Figure 1: Decomposition by rows parallelization strategy

 ISBN: 978-84-608-6082-2

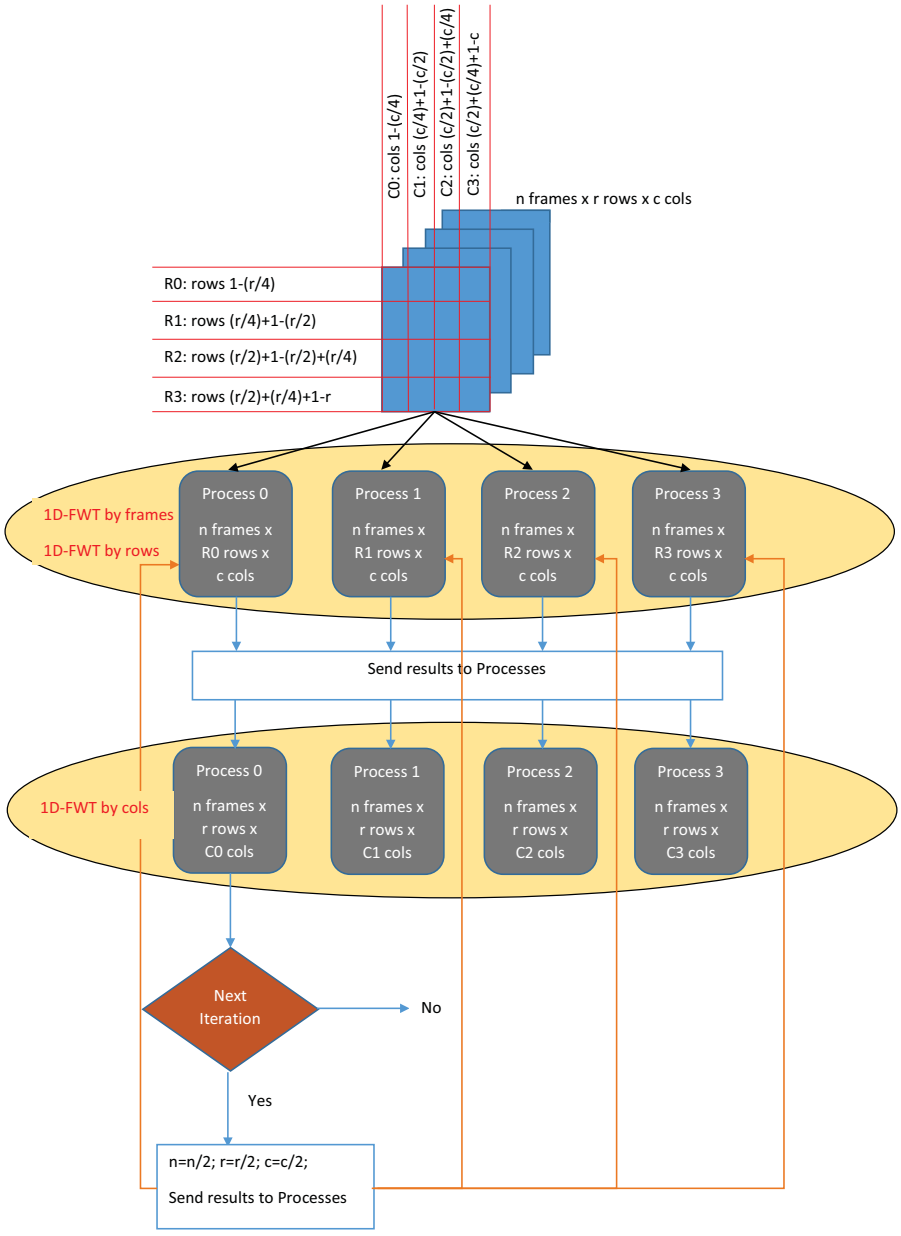
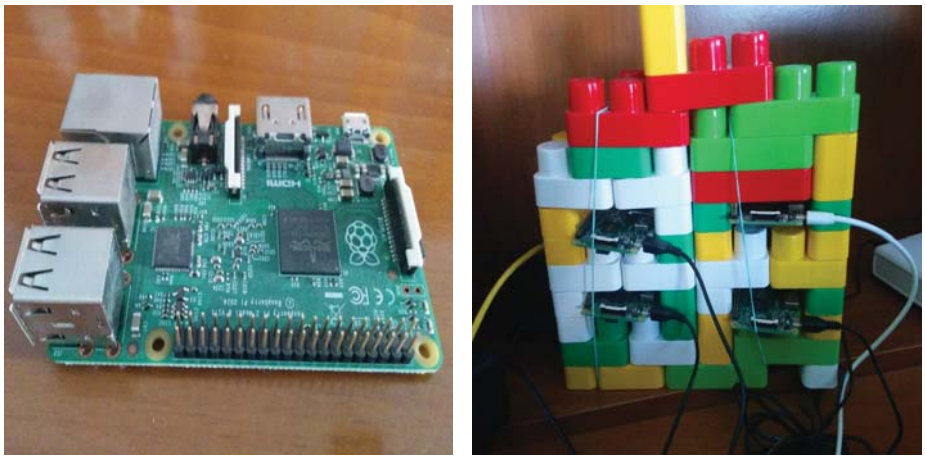


Figure 2: Decomposition by rows and columns parallelization strategy

 ISBN: 978-84-608-6082-2



(a) (b)

Figure 3: (a) Raspberry Pi 2 Model B. (b) Cluster of four Raspberry Pi 2 Model B nodes.

We have executed and measured the optimized 3D-FWT sequential version previously used in [7, 12] on a Raspberry Pi 2 for 2 iterations of the 3D fast wavelet transform. This will be baseline along the evaluation. Moreover, we have configured different parallel execution scenarios for the two parallel versions of the 3D-FWT explained before. We consider parallel executions with 2 and 4 MPI processes, running on the same Raspberry Pi 2 board or different boards. This allows us evaluating the impact that the limited bandwidth of the on-board LAN port has on the results.

Figure 4 shows the execution times (in seconds) for 2 iterations of the 3D-FWT on an input video containing 64 frames of $512x512$ pixels. From left to right, we first present the result obtained for the sequential version (`Sequential`), then we show the results for the decomposition by rows (`Rows`) and decomposition by rows and columns (`Rows&Cols`) parallelization strategies respectively. For each one of the two parallelization strategies, we consider 2 and 4 MPI processes (`2P` and `4P` respectively) running on one Raspberry Pi 2, and on 2 or 4 Raspberry Pi 2 (`2P-R` and `4P-R` respectively), having just 1 MPI process per board in each case.

From the results, we can see that the two proposed parallelization strategies of the 3D-FWT obtain noticeable speed-ups when executed on a single Raspberry Pi 2 with regard to the optimized sequential version. However, the executions on different Raspberry Pi 2 show negative outcomes from the performance point of view. What makes the differences is that in the first case all communications take place on the same board, and therefore, can be performed with low latency. Contrarily what happens when communications take place between several Raspberry Pi. In this case, all communications go through the low bandwidth on-board LAN ports, which ballast any performance advantages that parallel

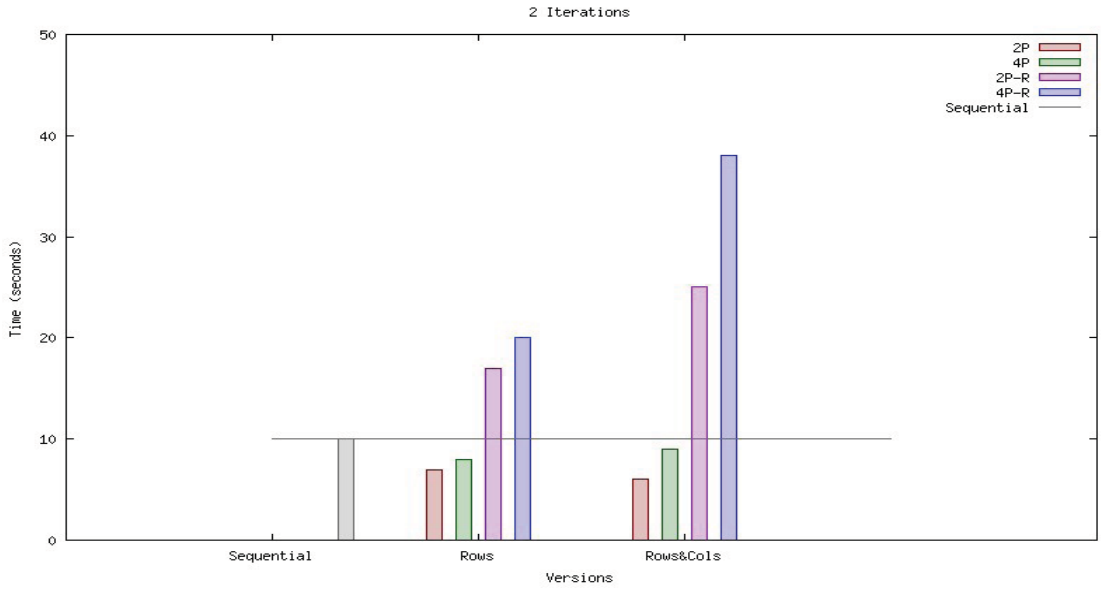


Figure 4: Execution times for 2 iterations of the 3D fast wavelet transform

execution could entail.

Taking a closer look to the results for one Raspeberry Pi board, we see that the speed-ups of the decomposition by rows parallelization strategy for 2 and 4 MPI processes are 1.56 and 1.34 respectively. Similarly, the decomposition by rows and columns parallelization strategy obtains speed-ups of 1.63 and 1.22 for 2 and 4 MPI processes respectively. Therefore, in both cases, fewer MPI processes, and therefore, lesser amount of communications among several processes, implies the best results. These two parallel versions do not scale due to the decreasing computation/communication ratio that they exhibit as the number of processes grows. Comparing the results for the two parallelization strategies, we observe that the increased potential parallelism exhibited by the decomposition by rows and columns strategy can only be exploited efficiently when the number of involved processes is small (2). In this case, the communication budget that it entails can be significantly amortized by the large amount of floating-point operations that each process performs. However, when the number of cores is 4, the larger communication requirements cannot be compensated, and lower speed-ups are obtained when compared with the decomposition by rows strategy.

In general, the speed-up results obtained within one Raspberry Pi keep in the expected range for a class of algorithms like the 3D-FWT, with low arithmetic intensity, pseudo-regular access patterns and intricate loop traversing. In these cases, maintaining communication requirements as low as possible is a key factor to obtain substantial profits.

5 Conclusions and future work

Current low-cost general-purpose single-board computing (SDC) devices are gaining increasing interests in research computing due to their very low cost/performance ratio. In this work we have analyzed the potential of a cluster built from Rasbperry Pi 2 SDCs as a viable platform for accelerating the execution of the 3D fast wavelet transform. To do so, we have implemented two parallelization strategies for the 3D fast wavelet transform using MPI. We found that although such cluster can constitute a very appealing solution from the cost point of view, it is unable to improve performance beyond what it is obtained when a single Raspberry Pi 2 is employed. That is, while noticeable speed-ups can be obtained when all MPI processes run on cores of a single Raspberry Pi 2 SDC, performance drops drastically when they spread to several boards. The reason for this is the limited bandwidth that the on-board LAN port (rated as 10/100 Fast Ethernet, and driven through the on-board USB 2.0 bus) can deliver, and that proves insufficient for the fine-grained, high-volume communication requirements of the two parallelization strategies.

Our plans for future work will delve in two orthogonal ways. On the one hand, there are several alternative coarse-grained parallelization strategies we would like to explore (for example, parallelizing by frames in a video sequence). Also, we are considering the possibility of using other SDCs different than Raspberry Pi for building the cluster. Particularly, we have found that Odroid C2 SDC may be an interesting alternative to consider since it provides, among other things, a Gigabit Ethernet LAN port which would help palliate the bandwidth limitations that our cluster configuration currently has.

Acknowledgements

This work was supported by the Spanish MINECO, as well as by European Commission FEDER funds, under grant TIN2015-66972-C5-3-R.

References

[1] M. Antonini and M. Barlaud. Image Coding Using Wavelet Transform. *IEEE Transactions on Image Processing*, 1(2):205–220, April 1992.

[2] S. Battista, F. Casalino, and C. Lande. MPEG-4: A Multimedia Standard for the Third Millenium, Part 1. IEEE Multimedia, October 1999.

[3] S. Battista, F. Casalino, and C. Lande. MPEG-4: A Multimedia Standard for the Third Millenium, Part 2. IEEE Multimedia, January 2000.

[4] G. Bernabé, R. Fernández, J. M. García, M. E. Acacio, and J. González. An Efficient Implementation of a 3D Wavelet Transform Based Encoder on Hyper-Threading Technology. *Journal of Parallel Computing*, 33(1):54–72, February 2007.

[5] G. Bernabé, J. M. García, and J. González. Reducing 3D Wavelet Transform Execution Time Using Blocking and the Streaming SIMD Extensions. *Journal of VLSI Signal Processing*, 41(2):209–223, 2005.

[6] G. Bernabé, J. González, J. M. García, and J. Duato. A New Lossy 3-D Wavelet Transform for High-Quality Compression of Medical Video. In *Proceedings of IEEE EMBS International Conference on Information Technology Applications in Biomedicine*, pages 226–231, November 2000.

[7] G. Bernabé, G. Guerrero, and J. Fernández. CUDA and OpenCL Implementations of 3D Fast Wavelet Transform. In *3rd IEEE Latin American Symposium on Circuits and Systems*, Playa del Carmen, Mexico, February 2012.

[8] Y. Chen and W. A. Pearlman. Three-Dimensional Subband Coding of Video Using the Zero-Tree Method. *Proc. of SPIE-Visual Communications and Image Processing*, pages 1302–1310, March 1996.

[9] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[10] J. Franco, G. Bernabé, J. Fernández, and M. E. Acacio. A Parallel Implementation of the 2D Wavelet Transform Using CUDA. In *17th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, Weimar, Germany, February 2009.

[11] J. Franco, G. Bernabé, J. Fernández, and M. Ujaldón. The 2D Wavelet Transform on Emerging Architectures: GPUs and Multicores. *Journal of Real-Time Image Processing. http://dx.doi.org/10.1007/s11554-011-0224-7*, (3):145–152, September 2012.

[12] J. Franco, G. Bernabé, J. Fernández, and M. Ujaldn. Parallel 3D Fast Wavelet Transform on manycore GPUs and multicore CPUs. In *10th International Conference on Computational Science*, Amsterdam, Netherlands, June 2010.

[13] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI, Second Edition*. MIT Press, 1999.

[14] A. Hague, G. Hastings, M. Klling, B. Croston, A. Oldknow, B. Lockwood, and C. Beale. *The Raspberry Pi Education Manual Version 1.0. Computing at School*. Creative Commons License, 2012.

[15] R. Heeks and A. Robinson. Ultra-low-cost computing and developing countries. *Communications of the ACM*, 56(8):22–24, August 2013.

[16] Y. Kim and W. A. Pearlman. Stripe-Based SPIHT Lossy Compression of Volumetric Medical Images for Low Memory Usage and Uniform Reconstruction Quality. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 2031–2034, 2000.

[17] A. S. Lewis and G. Knowles. Image Compression Using the 2D Wavelet Transform. *IEEE Transactions on Image Processing*, 1(2):244–256, April 1992.

[18] S. Mallat. A Theory for Multiresolution Signal Descomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[19] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek. An Overview of JPEG-2000. In *Proceedings of Data Compression Conference*, March 2000.

[20] P. Meerwald, R. Norcen, and A. Uhl. Cache Issues with JPEG2000 Wavelet Lifting. In *Proceedings of Visual Communications and Image Processing Conference*, pages 626–634, January 2002.

[21] P. Membrey and D. Hows. *Learn Raspberry Pi 2 with Linux and Windows 10 (2nd Edition)*. Apress, 2015.

[22] D. Santa-Cruz and T. Ebrahimi. A Study of JPEG 2000 Still Image Coding Versus Others Standards. In *Proceedings of X European Signal Processing Conference*, September 2000.

[23] N. Schot. Feasibility of raspberry pi 2 based micro data centers in big data applications. In *Proceedings of the 23rd Twente Student Conference on IT*, June 2015.

[24] A. Shahbahrami, B. Juurlink, and S. Vassiliadis. Improving the Memory Behavior of Vertical Filtering in the Discrete Wavelet Transform. In *Proceedings of ACM Conference in Computing Frontiers*, pages 253–260, September 2006.

[25] J. M. Shapiro. Embedded Image Coding Using Zerotrees of Wavelets Coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, December 1993.

[26] J. Tao, A. Shahbahrami, B. Juurlink, R. Buchty, W. Karl, and S. Vassiliadis. Optimizing Cache Performance of the Discrete Wavelet Transform Using a Visualization Tool. *Procs. of IEEE Intl. Symposium on Multimedia*, pages 153–160, December 2007.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Rational blends of two cones from square-root parameterized medial axis transforms

Michal Bizzarri[1] **and Miroslav Lávička**[1,2]

[1] *NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

[2] *Department of mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

emails: bizzarri@ntis.zcu.cz, lavicka@kma.zcu.cz

Abstract

In this paper we will continue in studying the problem of constructing a flexible blending canal surface smoothly joining two given circular cones, cf. [3]. We will use the main advantage of the contour method, cf. [5, 3], which allows us to generate in one step a whole family of rational canal surfaces having the same silhouette with respect to some prescribed vector. Involving a recently introduced class of RE curves (i.e., rational envelope curves, see [6]), we present a simple and efficient algorithm for constructing flexible blending canal surfaces employing only techniques for Hermite interpolation.

Key words: Canal surfaces; smooth transition; blending surface; contour curve; RE curve; Ferguson cubic.

1 Introduction

Canal surfaces, see e.g. [7], are envelopes of one parameter families of spheres in 3-space. They play a very important role in Computer-Aided Design as they can be used for constructing smooth transitions (blends) between two given surfaces, see [2] and references therein. For a constant radius function of generating spheres we speak about pipe surfaces. A well known non-trivial subclass of the canal surfaces are Dupin cyclides which are defined as the envelopes of all spheres touching three given spheres, see e.g. [8, 10, 12, 14].

It was proved in [13] that any canal surface with a rational spine curve and a rational radius function has a rational parameterization. An algorithm for computing rational

 ISBN: 978-84-608-6082-2

parameterizations of canal surfaces was presented in [11]. Let us emphasize that although the canal surfaces with rational spine curves and rational radii always possess exact rational parameterizations, approximate parameterization techniques are also investigated in connection with them, see e.g. [4]. This is caused by the computational difficulty of decomposing a rational function into a sum of two squares (SOS) over reals, which is a necessary part of the parameterization algorithm from [11]. Moreover, we would like to recall that that there exist rational canal surfaces with square-root parameterized medial axis transform (which is a curve i 4-space obtained from the spine curve when the radius function is added as the fourth coordinate).

We will continue in the study of the problem to construct a flexible blending canal surface smoothly joining two given cones, cf. [3]. This is especially useful when it is required that the constructed joint shall bypass some given bounded obstacle(s). We will use the main advantage of the contour method, cf. [5, 3] enabling us to generate in one step a whole family of rational canal surfaces having the same silhouette with respect to some prescribed vector. Involving a recently introduced class of RE curves (i.e., rational envelope curves, see [6]), we present a simple and efficient algorithm for constructing flexible blending canal surfaces employing only techniques for Hermite interpolation of curves.

The remainder of this paper is organized as follows. The next section summarizes some fundamental facts about canal surfaces and rational envelope (RE) curves. In Section 3, a simple method for constructing transition canal surfaces using rational contour curves with respect to a prescribed viewpoint is presented. Section 4 is devoted to a technique for constructing a flexible blending canal surfaces which can be used when bypassing a given obstacle is needed. Finally, in Section 5 we conclude the paper.

2 Preliminaries

Consider a one-parameter set of spheres

$$F(t): \ |(\mathbf{m}(t)-\mathbf{x})|^2 - r^2(t) = 0, \quad t \in I \subseteq \mathbb{R}, \tag{1}$$

where $\mathbf{x} = (x, y, z)$. The envelope $\mathcal{S}$ of $F(t)$ is called a *canal surface*, the curve $\mathbf{m}(t)$ tracing the centres of the spheres $F(t)$ its *spine curve* and the function $r(t)$ describing the radii of $F(t)$ its *radius function*. By appending the corresponding sphere radii $r(t)$ to the points of the spine curve $\mathbf{m}(t)$ we obtain the *medial axis transform* (MAT for short). For the sake of clarity, we identify the canal surface $\mathcal{S}$ with its medial axis transform $\big(\mathbf{m}(t), r(t)\big) \subset \mathbb{R}^{3,1}$, where $\mathbb{R}^{3,1}$ is the 4-dimensional *Minkowski space*. In what follows we omit the dependence on parameter t whenever no confusion is likely to arise.

The rational parametrization of a canal surface $\mathcal{S}$ can be computed by rotating any

rational curve $\mathbf{c}$ on $\mathcal{S}$ around the tangents of the spine curve $\mathbf{m}$, i.e,

$$\mathbf{s}(t,u) = \mathbf{m}(t) + \frac{(\varrho(u) + \mathbf{m}'(t)) \star (\mathbf{c}(t) - \mathbf{m}(t)) \star (\varrho(u) - \mathbf{m}'(t))}{(\varrho(u) + \mathbf{m}'(t)) \star (\varrho(u) - \mathbf{m}'(t))}, \tag{2}$$

where $\varrho(u)$ is any arbitrary rational function, the sums $\varrho(u) \pm \mathbf{m}'(t)$ of scalars and vectors are considered as quaternions, and $\star$ is the operation of quaternion multiplication, see [1, 9] for more details about quaternions.

In this paper we employ canal surfaces for constructing rational smooth transition surface between two prescribed parts of cones $\mathcal{S}_1$ and $\mathcal{S}_2$ (or cylinders in special situations). The cones (as special instances of canal surfaces) can be described by their MATs, which are straight lines now. We assume that these lines are determined by a point and a direction vector in $\mathbb{R}^{3,1}$, in particular

$$\overline{\mathbf{s}}_i = (\mathbf{s}_i, r_i) \quad \text{and} \quad \overline{\mathbf{u}}_i = (\mathbf{u}_i, \tau_i), \qquad i = 0, 1. \tag{3}$$

The construction of a joining canal surface is based on finding some suitable rational spine curve $\mathbf{m}$ of this considered blend and some rational curve $\mathbf{c}$ on it. Then we can compute the rational parametrization by (2). Let us emphasize that the presented approach can be considered as well when the input shapes are general canal surfaces by considering the corresponding tangent cones at the end circles.

Recently introduced *rational envelope* (*RE*) *curves*, cf. [6], affort a way how the curves $\mathbf{m}$ and $\mathbf{c}$ can be find symbolically while using only Hermite (curve) interpolation techniques for C^1/G^1 data, whereas the classical blending algorithms (which are based on computing rational parametrization of canal surfaces from their MATs) use SOS (sum of squares) decomposition, which involves some numerical steps, or on more complicated interpolations with Minkowski Pythagorean hodograph curves, see e.g. [11, 5], .

Let us recall that RE curves are defined as the curves in $\mathbb{R}^{2,1}$ (not necessarily rational) with the distinguished property that when considered as MATs they lead to the associated rational envelopes of circles in plane $\mathbb{R}^2$. It was shown in [6] that any RE curve $\overline{\mathbf{y}}$ in $\mathbb{R}^{2,1}$ can be constructed starting from a rational curve $\mathbf{x}$ lying in the coordinate plane $xy: z = 0$ and a rational function f in the form

$$\overline{\mathbf{y}} = \left(\mathbf{x} + f\,\mathbf{x}'^{\perp}, f|\mathbf{x}'|\right), \tag{4}$$

where $\mathbf{v}^{\perp}$ denotes the orthogonal vector to $\mathbf{v}$.

3 Rational blending canal surfaces with rational contour curves

In this section we discuss a simple method for constructing smooth joint canal surface (between two prescribed cones) based on using rational *contour curves*, i.e., curves on surface

 ISBN: 978-84-608-6082-2

corresponding to the silhouettes of the surface in plane with respect to a prescribed viewpoint $\mathbf{v}$. Contour curves (in the blending process) allows us to obtain in one step the whole family of transition surfaces with the same silhouette in the projection plane

$$\mathcal{P}_{\mathbf{v}} : \ \mathbf{x} \cdot \mathbf{v} = 0. \tag{5}$$

This is especially useful when some object must be bypassed or avoided, see Section 4. Using RE curves we will be able to construct a family of blending canal surfaces (based on any arbitrary interpolation with rational curves) with rational contour curves (and thereby rational silhouettes) with respect to an arbitrary, intentionally chosen, viewpoint $\mathbf{v}$.

Now, we describe the particular steps of the blending algorithm in more detail. The algorithm starts with two cones $\mathcal{S}_1$ and $\mathcal{S}_2$ given by (3). We choose a (suitable) unit vector $\mathbf{v}$, cf. Section 4, and transform data (3) to the form

$$\overline{\mathbf{p}}_i = (\pi_{\mathbf{v}}(\mathbf{s}_i), r_i) \quad \text{and} \quad \overline{\mathbf{t}}_i = (\pi_{\mathbf{v}}(\mathbf{u}_i), \tau_i)\,, \tag{6}$$

where $\pi_{\mathbf{v}} : \mathbb{R}^3 \to \mathcal{P}_{\mathbf{v}}$ is the orthogonal projection realized as follows

$$\pi_{\mathbf{v}}(\mathbf{x}) = \mathbf{x} - (\mathbf{x} \cdot \mathbf{v})\mathbf{v}. \tag{7}$$

From the envelope formula, cf. [6], the associated points $\mathbf{q}_i \in \mathcal{P}_{\mathbf{v}}$ on the corresponding envelope curve will be of the form

$$\mathbf{q}_i = \mathbf{p}_i - r_i \frac{\tau_i\, \mathbf{t}_i + (\mathbf{t}_i \times \mathbf{v}) \sqrt{|\mathbf{t}_i|^2 - \tau_i^2}}{|\mathbf{t}_i|^2}. \tag{8}$$

The corresponding vectors $\mathbf{w}_i$ also lie in $\mathcal{P}_{\mathbf{v}}$ and must be perpendicular to the vectors $\mathbf{p}_i - \mathbf{q}_i$, i.e., we arrive at

$$\mathbf{w}_i = \alpha_i\, (\mathbf{p}_i - \mathbf{q}_i) \times \mathbf{v}. \tag{9}$$

The parameters α_i, corresponding to the magnitudes of $\mathbf{w}_i$, are free parameters and can be chosen to modify the resulting shape. Next, we interpolate $\mathbf{q}_i$ and $\mathbf{w}_i$ in the plane $\mathcal{P}_{\mathbf{v}}$ by a suitable Ferguson cubic $\mathbf{x}(t)$, $t \in [0, 1]$ and compute the MAT $\overline{\mathbf{y}}$, i.e.,

$$\overline{\mathbf{y}} = \left(\mathbf{x} + (\mathbf{x}' \times \mathbf{v}) f, f\, |\mathbf{x}'|\right). \tag{10}$$

The function f can be constructed as a one dimensional Ferguson cubic interpolating

$$f_i = \frac{r_i}{|\mathbf{w}_i|}, \quad f_i' = -\frac{\mathbf{t}_i \cdot (f_i\, \mathbf{x}''(i) + (\mathbf{w}_i \times \mathbf{v}))}{\mathbf{t}_i \cdot \mathbf{w}_i}, \tag{11}$$

which follows from the interpolation conditions

$$\overline{\mathbf{y}}(i) = \overline{\mathbf{p}}_i \quad \text{and} \quad \overline{\mathbf{y}}'(i) = \beta_i\, \overline{\mathbf{t}}_i. \tag{12}$$

Moreover

$$\beta_i = \frac{\mathbf{w}_i \cdot \mathbf{w}_i - f_i\, \mathbf{x}''(i) \cdot (\mathbf{w}_i \times \mathbf{v})}{\mathbf{t}_i \cdot \mathbf{w}_i}. \tag{13}$$

To lift $\mathbf{y}$ from $\mathcal{P}_{\mathbf{v}}$ we interpolate data

$$\mathbf{s}_i \quad \text{and} \quad \beta_i\, \mathbf{u}_i \tag{14}$$

by a Ferguson cubic $\mathbf{h}(t)$ and construct the spine curve of the corresponding canal surface in the form

$$\mathbf{m} = \mathbf{y} + (\mathbf{h} \cdot \mathbf{v})\mathbf{v}, \qquad t \in [0,1]. \tag{15}$$

Additionally, the curve

$$\mathbf{c}_{\mathbf{v}} = \mathbf{x} + (\mathbf{h} \cdot \mathbf{v})\mathbf{v}, \qquad t \in [0,1] \tag{16}$$

is, by construction, a rational contour curve of the constructed canal surface with respect to the vector $\mathbf{v}$. Finally, rotating $\mathbf{c}_{\mathbf{v}}$ around the tangents of $\mathbf{m}$ yields a rational parameterization of the seeking blending canal surface, cf. (2).

4 Flexible blending canal surfaces

In this section we present a method for constructing a rational canal surface smoothly joining two cones which bypasses a given obstacle. For the sake of simplicity, we consider that the obstacle is inscribed into a sphere, however the presented blending method can be easily generalized for any type of obstacle.

The first step of the algorithm is to choose a suitable direction $\mathbf{v}$ and using the approach introduced in Section 3 to compute the rational spine curve $\mathbf{m}$ and the rational contour curve $\mathbf{c}_{\mathbf{v}}$ with respect to the vector $\mathbf{v}$. Then if the canal surface determined by $\mathbf{m}$ and $\mathbf{c}_{\mathbf{v}}$ intersects the given obstacle we create a new blending canal surface described by a new spine and contour curve of the form

$$\mathbf{m}^* = \mathbf{m} + \varrho\, \mathbf{v}, \quad \mathbf{c}^*_{\mathbf{v}} = \mathbf{c}_{\mathbf{v}} + \varrho\, \mathbf{v}, \tag{17}$$

where $\varrho(t)$ is a rational function, called the *distance function*, continuous on $(0,1)$ and simultaneously satisfying the constrains

$$\varrho(0) = 0, \quad \varrho'(0) = 0, \quad \varrho(1) = 0, \quad \varrho'(1) = 0. \tag{18}$$

The distance function ϱ enables to modify the resulting joint in the $\mathbf{v}$-direction since the original and the new transition surfaces possess the same projection with respect to $\mathbf{v}$.

Now, we show how the direction $\mathbf{v}$ in which the constructed canal surface possesses the rational contour curve (and thereby a direction in which it will be adapted) can be determined. Consider two cones given by (3) and a spherical obstacle with the center $\mathbf{o}$

 ISBN: 978-84-608-6082-2

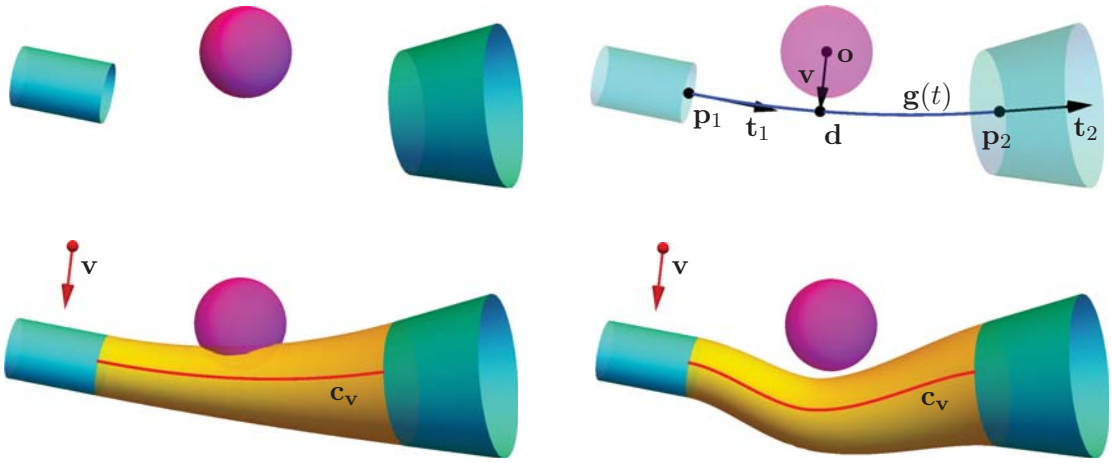


Figure 1: Construction of a blending canal surface between two cones avoiding a prescribed obstace (sphere) from Example 4.1.

and the radius $r_{\mathcal{S}}$. First, we interpolate data $\mathbf{s}_i, \mathbf{u}_i \in \mathbb{R}^3$ by a Ferguson cubic $\mathbf{g}(t)$ which approximately simulates the shape of a prospective spine curve of the blend. Then we compute the direction in which the distance of $\mathbf{g}$ and the point $\mathbf{o}$ can be suitably adapted. In particular by minimizing

$$\varphi(t) = |\mathbf{g}(t) - \mathbf{o}|^2, \quad t \in [0,1], \tag{19}$$

we compute the parameter $\tau \in (0,1)$ and therefore the point $\mathbf{d} = \mathbf{g}(\tau)$ such that $|\mathbf{d} - \mathbf{o}|$ is minimal. Finally we choose the vector $\mathbf{v}$ equal to:

$$\mathbf{v} = \frac{\mathbf{o} - \mathbf{d}}{|\mathbf{o} - \mathbf{d}|}. \tag{20}$$

Next, we proceed as in Section 3, i.e., we compute the spine curve $\mathbf{m}$ and contour curve $\mathbf{c_v}$ and apply a suitable distance function f, cf. (17) chosen in such a way that the function ψ, responsible for the distance of the blending surface and the obstacle, is positive.

$$\psi = |\mathbf{m}^*(t) - \mathbf{o}|^2 - (r + r_{\mathcal{S}})^2. \tag{21}$$

Example 4.1 Consider two cones given by

$$\overline{\mathbf{s}}_0 = (0,0,0,1), \;\; \overline{\mathbf{s}}_1 = (10,1,0,0); \qquad \overline{\mathbf{u}}_0 = (10,1,2,2), \;\; \overline{\mathbf{u}}_1 = (10,1,4,2), \tag{22}$$

and a spherical obstacle with the center $\mathbf{o} = (4, 2, 3/2)$ and the radius $r_{\mathcal{S}} = 3/2$, see Fig. 1 (top, left). First, we find a suitable vector $\mathbf{v}$ by the method described above, see Fig. 1 (top, right). Next we compute the corresponding blending surface, see Fig. 1 (bottom, left), which however intersects the prescribed obstacle. Thus we choose a distance function

$$\varrho(t) = \begin{cases} 4(3-4t)t^2, & t \in [0, 1/2], \\ 4(t-1)^2(4t-1), & t \in [1/2, 1], \end{cases} \tag{23}$$

such that function (21) is positive. Finally we compute a new spine and contour curves, cf. (17), and afterwards the rational parameterization of the joint, see. Fig. 1 (bottom, right).

5 Conclusion

In this manuscript we continued with studying the contour method for the construction of smooth transition canal surface between prescribed cones. We focused on one particular application of recently introduced rational envelope curves considered here as projections of MATs of canal surfaces. Using these curves significantly simplified the known approaches and enabled to solve situations which were unsolvable before. We focused mainly on constructing suitable flexible blends sharing the same silhouettes, which can be easily used in the practical problem of avoiding obstacles. The designed method was presented on one particular example.

Acknowledgements

The authors are supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

[1] S. L. Altmann, *Rotations, quaternions, and double groups*, Dover Publications, 2005.

[2] B. Bastl, B. Jüttler, M. Lávička, and T. Schulz, *Blends of canal surfaces from polyhedral medial surface transform representations*, Computer-Aided Design Design, 43 (2011), pp. 1477–1484.

[3] M. Bizzarri, M. Lávička, and J. Vršek, *Canal surfaces with rational contour curves and blends bypassing the obstacles*, Computer-Aided Design, 64 (2015), pp. 55 – 67.

[4] M. Bizzarri and M. Lávička, *A symbolic-numerical method for computing approximate parameterizations of canal surfaces*, Computer-Aided Design, 44 (2012), pp. 846 – 857.

[5] ——, *Parameterizing rational offset canal surfaces via rational contour curves*, Computer-Aided Design, 45 (2013), pp. 342 – 350. Solid and Physical Modeling 2012.

[6] M. Bizzarri, M. Lávička, and J. Kosinka, *Medial axis transforms yielding rational envelopes*, Submitted to Computer Aided Geometric Design (under revision), (2015).

[7] M. Dohm and S. Zube, *The implicit equation of a canal surface*, Journal of Symbolic Computation, 44 (2009), pp. 111–130.

[8] D. Dutta, R. R. Martin, and M. J. Pratt, *Cyclides in surface and solid modeling*, IEEE Computer Graphics and its Applications, 13 (1993), pp. 53–59.

[9] R. Goldman, *Understanding quaternions*, Graphical Models, 73 (2011), pp. 21–49.

[10] R. Krasauskas and C. Mäurer, *Studying cyclides with Laguerre geometry*, Computer Aided Geometric Design, 17 (2000), pp. 101–126.

[11] G. Landsmann, J. Schicho, and F. Winkler, *The parametrization of canal surfaces and the decomposition of polynomials into a sum of two squares*, Journal of Symbolic Computation, 32 (2001), pp. 119 – 132.

[12] M. Lávička and J. Vršek, *On the representation of Dupin cyclides in Lie sphere geometry with applications*, Journal for Geometry and Graphics, 13 (2009), pp. 145–162.

[13] M. Peternell and H. Pottmann, *Computing rational parametrizations of canal surfaces*, J. Symb. Comput., 23 (1997), pp. 255–266.

[14] H. Pottmann and M. Peternell, *Applications of Laguerre geometry in CAGD*, Computer Aided Geometric Design, 15 (1998), pp. 165–186.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Approximating Support Function at Inflection Points for CNC Manufacturing

Eva Blažková[1] **and Zbyněk Šír**[1]

[1] *Faculty of Mathematics and Physics, Charles University in Prague*

emails: eblazkova@karlin.mff.cuni.cz, sir@karlin.mff.cuni.cz

Abstract

We study the behavior of the support function in the neighborhood of a curve inflection. The gauss map at the inflection point is not regular and in the neighborhood is typically not injective. The support function is thus not regular and typically multivalued. We describe this function using an implicit algebraic equation and the Puiseux series of its branches. We show the correspondence between the degree of the approximation of the primary curve (using Taylor series) and the degree of the approximation of the support function (using Puiseux series). Based on this results we are able to approximate curve with inflections by curves with a simple support function which consequently possess rational offsets.

Key words: support function, critical points, inflections approximation, Puiseux series

1 Introduction

The support function representation describes a curve as the envelope of its tangent lines, where the distance between the tangent line and the origin is specified by a function of the unit normal vector. This representation is one of the classical tools in the field of convex geometry [7]. In this representation offsetting and convolution of curves correspond to simple algebraic operations of the corresponding support functions. In addition, it provides a computationally simple way to extract curvature information [4]. Applications of this representation to problems from Computer Aided Design were foreseen in the classical paper [9] and developed in several recent publications, see e.g., [1, 2, 3, 5, 6, 8, 10, 11].

The main importance of this representation consists in the possibility of an efficient simultaneous approximation of a shape and its offsets or more generally its convolution with

 ISBN: 978-84-608-6082-2

some other shape. This fact can be applied in the context of the CNC manufacturing, where the tool center does not follow the produced object boundary, but rather the convolution with the tool shape. Using the support function representation allows to avoid complicated approximation techniques in the offset/convolution description.

So far the inflections were excluded from the description using support function. The main reason is that the normal turns back at inflections and the support function must be considered as multivalued in its neighborhood. The purpose of this paper is to analyze the behavior of the support function in the neighborhood of a curve inflection and use it to design an applicable interpolation scheme.

The reminder of the paper is organized as follows. In Section 2 we give the basic definitions and properties of the explicit and implicit support function. In Section 3 we revisit the topic of the C^1 Hermite interpolation and show that the approximation order drops from 4 to 3 when an inflection occurs. In Section 4 we analyze the behavior of the support function at inflections and design an interpolation strategy with the approximation order 4. Eventually, we conclude the paper.

2 Support function representation of curves

For a parametrically or implicitly given planar curve $\mathcal{C}$ we define the support function h as a (possibly multivalued) function defined on (a subset of) the unit circle

$$h : \mathbb{S}^1 \supset U \to \mathbb{R}^1$$

which to a unit normal $\mathbf{n} = (n_1, n_2)$ associates the oriented distance(s) from the origin $[0, 0]$ to the corresponding tangent line(s) of the curve.

As shown e.g. in [11] we can recover the curve $\mathcal{C}$ from $h(\mathbf{n})$ as the envelope of the system of the tangent lines $\{\mathbf{n} \cdot \mathbf{x} - h(\mathbf{n}) = 0 : \mathbf{n} \in U\}$. This envelope is locally parameterized via the formula

$$\mathbf{c}_h(\mathbf{n}) = h(\mathbf{n})\mathbf{n} + \nabla_{\mathbb{S}_1} h(\phi) = h(\mathbf{n})\mathbf{n}(\phi) + \dot{h}(\phi)\dot{\mathbf{n}}(\phi), \tag{1}$$

where $\nabla_{\mathbb{S}_1}$ denotes the intrinsic gradient with respect to the unit circle, which is alternatively expressed using the following arc length parameterization of $\mathbb{S}_1$

$$\mathbf{n}(\phi) = (\cos(\phi), \sin(\phi)), \qquad \dot{\mathbf{n}}(\phi) = (-\sin(\phi), \cos(\phi)). \tag{2}$$

It is also possible to parameterize bi-rationally the unit circle

$$\mathbf{n} = [n_1, n_2] = \left[\frac{1 - s^2}{1 + s^s}, \frac{2s}{1 + s^2}\right] \tag{3}$$

and obtain a kind of the affine version of the support function together with the rational parameterization of the curve $h(s)$ and $\mathbf{c}_h(s)$. Note, that the inversion formulas for changing the variable is $s = \frac{n_2}{1+n_1} = \arctan(\phi/2)$.

 ISBN: 978-84-608-6082-2

As there are often many tangent lines with the same normal it is globally not always possible to obtain an explicit expression of h but rather an implicit one, which is closely related to the notion of dual curve consisting of the tangent lines of $\mathcal{C}$.

For example in the case of a polynomial $\mathcal{C}$ given as the zero set of a polynomial $f(x, y)$, the equation of the dual curve $D(h, \mathbf{n}) = 0$ can be computed by eliminating x and y from the following system of equations:

$$f(x, y) = 0, \quad \mathbf{n}^{\perp} \cdot \nabla f = 0, \quad \mathbf{n} \cdot [x, y] = h. \tag{4}$$

The dual equation $D(h, \mathbf{n}) = 0$ together with the algebraic constraint $n_1^2 + n_2^2 = 1$ is called the implicit definition of the support function h or simply *the implicit support function*. The implicit support function is obviously a kind of dual representation which takes into account the Euclidean metric. Using the parameterization (3) we get the affine implicit support function $D(h, s) = 0$.

Support function representation has many nice properties and in particular simplifies the offset and convolution computation cf. [9, 11]. In particular if the support function $h(s)$ is rational, not only the resulting parameterization $\mathbf{c}(s)$ will be rational, but it will also provide rational offsets. In fact the offsetting corresponds to adding a constant to the support function. More generally the convolution of curves is obtained via adding the support functions. Rotation and translation is obtained by simple modifications of the support. Also the curvature of the curve at the corresponding point can be expressed by the simple linear formula

$$\kappa = -\frac{1}{h + \ddot{h}} \tag{5}$$

and the points where $h + \ddot{h} = 0$ correspond to cusps.

Example 2.1 Consider the Tschirnhausen cubic $f(x, y) = x^3 - 9x^2 + 27y^2$. By simultaneous solving $f(x, y) = 0$, $\nabla f \cdot (n_2, -n_1) = 0$ and $n_1 x + n_2 y = h$ we get its support function

$$h(n_1, n_2) = \frac{2 - 2n_1 + 8n_1 n_2^2}{n_2^2}.$$

The geometric meaning of the support function is clear from Figure 1.

Using the arc length parametrization (2) of the unit circle, we get

$$h(\phi) = \frac{2 - 2\cos(\phi) + 8\cos(\phi)\sin^2(\phi)}{\sin^2(\phi)}.$$

Using the bi-rational parametrization (3) of the unit circle, we get

$$h(s) = \frac{\left(3 - s^2\right)^2}{s^2 + 1}.$$

 ISBN: 978-84-608-6082-2

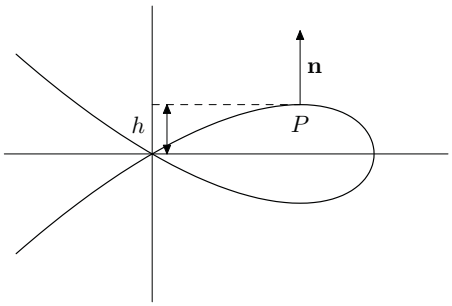


Figure 1: Support function of the Tschirnhausen cubic at a given point P.

The dual equation of Tschirnhausen cubic is

$$D(h, \mathbf{n}) = 2 - 2n_1 + 8n_1n_2^2 - hn_2^2.$$

This equation with the constraint $n_1^2 + n_2^2 = 1$ defines the implicit support function.

3 C^1 Hermite interpolation using the support function

In this section we will show how to interpolate C^1 Hermite data at two points via the support function. We will see that this construction leas to a linear system of equations on the coefficients of suitably chosen support function. We will also show on an example that the approximation degree of this construction is lower when an inflection occurs.

Let $[x_0, y_0]$, $[x_1, y_1]$ be two end-points of a segment of the curve $\mathcal{C}$ and ϕ_0, ϕ_1 the angles of the corresponding oriented normals. More precisely, we are given the first order geometric data

$$([x_0, y_0], \mathbf{n}(\phi_0)) \text{ and } ([x_1, y_1], \mathbf{n}(\phi_1)) . \tag{6}$$

We will interpolate this G^1 boundary data using the support function. More precisely we look for a function h defined over the interval $[\phi_0, \phi_1]$, so that the curve segment obtained via the formula (1) interpolates the boundary points.

From the form of (1) it is clear, that the position of a point provides values of the support function and its derivative. The G^1 interpolation by a smooth segment is thus transformed into the standard problem of the functional Hermite interpolation of the first order at the two values ϕ_0 and ϕ_1. Natural choices of the basis functions are trigonometric polynomials [11], trigonometric polynomial with rational non-integer arguments [3] and rational trigonometric functions [6]. But in the view of the subsequent higher order interpolation at the inflection

we will consider the support function dependent on the rational variable s defined by (3). More precisely we will use the support function in the form

$$h(s) = a_0 + a_1 s + a_2 s^2 + a_3 s^3. \tag{7}$$

To interpolate (6) we first determine explicitly the corresponding values

$$s_0 = \arctan(\phi_0/2), \quad s_1 = \arctan(\phi_1/2)$$

and then we have to solve the interpolating equations

$$\mathbf{c}_h(s_0) = [x_0, y_0], \quad \mathbf{c}_h(s_1) = [x_1, y_1].$$

In fact in these last equations the coefficients a_i appears linearly due to the form of (1) and we find them solving four linear equations in four variables.

It is a great advantage of the support function, that its approximation error will translate to the identical behavior of the Hausdorff distance of corresponding segments.

Proposition 3.1 *Let h, g be two support functions defined on the interval $U = [s_0, s_1]$, such that*

$$g(s_i) = h(s_i), \qquad g'(s_i) = h'(s_i), \qquad i \in 0, 1\ .$$

Suppose, that the corresponding curve segments $\mathbf{c}_h$, $\mathbf{c}_g$ are cusp-free on U. Then their Hausdorff distance is equal to the error in support functions

$$||\mathbf{c}_h - \mathbf{c}_g||_H = ||h - g||_\infty = \max_{s \in [s_0, s_1]} |h(s) - g(s)|. \tag{8}$$

Proof: Due to boundary conditions and absence of singular points (cusps), the Hausdorff distance is realized by a common normal line to both curve segments. The distance of the points on this line is equal to the absolute value of the difference of the support functions. For a more formal proof see [11, Proposition 14]. □

For this reason, to estimate the Hausdorff it is not necessary to perform a sampling from both curves (requiring a quadratic complexity), but rather sample linearly the values of $s \in [s_0, s_1]$. If the error is too big, we subdivide the curve segment in more parts. The improvement of the error depending on the number of parts is called approximation order of the interpolation. We will investigate this important value on the following example.

Example 3.2 Let us consider the parametric curve

$$\mathbf{c}(t) = [(-5 + 8t - 6t^3 - 6t^4 - 6t^5)/10, (5 - 3t - 4t^3 - 4t^4 - 4t^5)/5],$$

see Fig. 2. Note that it has an inflection at $\mathbf{c}(0) = [-1/2, 1]$.

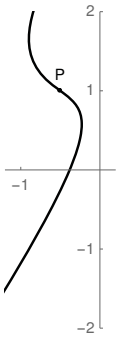


Figure 2: Parametric curve with an ordinary inflection.

Let us first consider the segment of $\mathbf{c}(t)$ for $t \in [0.25, 1.25]$ which does not contain the inflection. We first perform the C^1 Hermite interpolation on the whole segment, obtaining the error $9.28701 \cdot 10^{-3}$. Then we divide in the middle and perform the interpolation of each segment. The global error drops to $2.15398 \cdot 10^{-3}$. We iterate this procedure and show the results in Table 1 (left), where n indicates the number of splitting segments. The third column gives the ratio of the previous and new error. We see that it tends to $16 = 2^4$, indicating the approximation order of 4.

Table 1: Errors for C^1 Hermite interpolation.

segment without inflection			segment with inflection		
n	error	error improvement	n	error	error improvement
1	$9.28701 \cdot 10^{-3}$		1	$2.51863 \cdot 10^{-3}$	
2	$2.15398 \cdot 10^{-3}$	4.31154	2	$2.15398 \cdot 10^{-3}$	6.97205
4	$1.43050 \cdot 10^{-4}$	15.05760	4	$2.32844 \cdot 10^{-4}$	10.8168
8	$1.33136 \cdot 10^{-5}$	10.74464	8	$2.51388 \cdot 10^{-5}$	9.26235
16	$9.27833 \cdot 10^{-7}$	14.34912	16	$2.93471 \cdot 10^{-6}$	8.56601
32	$1.86660 \cdot 10^{-7}$	14.97071	32	$3.55065 \cdot 10^{-7}$	8.26528
64	$1.18982 \cdot 10^{-8}$	15.68805	64	$4.36787 \cdot 10^{-8}$	8.12903
128	$7.50893 \cdot 10^{-10}$	15.84542	128	$5.41676 \cdot 10^{-9}$	8.06361
256	$4.71333 \cdot 10^{-11}$	15.93125	256	$6.74433 \cdot 10^{-10}$	8.03158
512	$2.94720 \cdot 10^{-12}$	15.99257	512	$8.41386 \cdot 10^{-11}$	8.01573

Next, let us do the same for the segment of $\mathbf{c}(t)$ for $t \in [0, 1]$, i.e. the segment with one inflection point. The results are displayed in Table 1 (right), which indicates in this case

the approximation order of 3.

This less efficient behavior at the inflection points confirms the results of [10], where some other kind of support function was used.

4 Support function at curve inflections

The behavior of the support function at inflections was never analyzed before. The problem is, that the support function is necessarily multivalued at the neighborhood of an inflection and we have to use some techniques related to the algebraic functions and Puiseux expansions.

Proposition 4.1 *Let $\mathcal{C}$ be a curve with an inflection with the normal $\mathbf{n}(s_0)$. Then the support function has the Puiseux expansion at the point $s = s_0$ of the form*

$$s = s_0 \pm r^2, \qquad h = b_0 + b_2 r^2 + b_3 r^3 + \mathcal{O}(r^4) \tag{9}$$

Proof: Let us consider the curve $\mathcal{C}$ with inflection at the point $[0,0] \in \mathcal{C}$ and the corresponding normal $[1,0]$. Supposing that the inflection is of the first order (the tangent has triple intersection with the curve), we can locally parametrize the curve by a parameterization with the Taylor expansion

$$\mathbf{c}(t) = [y_3 t^3 + \mathcal{O}(t^4), t]. \tag{10}$$

By an explicit series computation we obtain that the variation of the normal expressed in the parameterization (3) has the expansion

$$s = -\frac{3}{2} y_3 t^2 + \mathcal{O}(t^3) \tag{11}$$

and the corresponding distance of the tangent to the origin (support function)

$$h = -2 y_3 t^3 + \mathcal{O}(t^4).$$

Composing with the inverse series to (11) we get the fractional Puiseux expansion for $h(s)$

$$h(s) = \frac{4}{3}\sqrt{\frac{2}{3|y_3|}}\, s^{3/2} + \mathcal{O}(s^2).$$

Relaxing to the Puiseux series with integer coefficients and allowing for the rotation and translation of the curve, we get the series (9), where s_0 is given by the position of the normal at the inflection, $b_0 = y_0$ and $b_2 = 2x_0$, where $[x_0, y_0]$ is the position of the inflection. $\square$.

From the previous proof we also obtain the following corollary

 ISBN: 978-84-608-6082-2

Corollary 4.2 *Suitable choice of the parameters in the support function of the form* (9) *will provide approximation of the curve inflection point up to the third order.*

This result suggests a modified strategy for the C^1 Hermite interpolation at the inflection points. We will use the support function in the form

$$s = s_0 \pm r^2, \qquad h = b_0 + b_2 r^2 + b_3 r^3 + b_4 r^4 + b_5 r^5 \tag{12}$$

and use the first three parameters b_0, b_2, b_3 to interpolate the inflection $\mathbf{c}(t_0)$ and the two remaining parameters b_4, b_5 will be set to interpolate the second (non-inflection) point $\mathbf{c}(t_1)$. This construction can be easily implemented and we obtained the results displayed at Table 2 (left).

Table 2: Errors for improved C^1 Hermite interpolation at inflectios.

division at half			division at the ratio $(l : l^{5/4})$		
n	error	error improvement	n	error	error improvement
1	$1.39501 \cdot 10^{-3}$		1	$4.50631 \cdot 10^{-2}$	
2	$1.52748 \cdot 10^{-4}$	9.13278	2	$2.79362 \cdot 10^{-3}$	16.1308
4	$1.23054 \cdot 10^{-5}$	12.4130	4	$1.91388 \cdot 10^{-4}$	14.5966
8	$1.19461 \cdot 10^{-6}$	10.3008	8	$1.24158 \cdot 10^{-5}$	15.4149
16	$1.32617 \cdot 10^{-7}$	9.00800	16	$8.43312 \cdot 10^{-7}$	14.7227
32	$1.56662 \cdot 10^{-8}$	8.46516	32	$5.27206 \cdot 10^{-8}$	15.9959
64	$1.90501 \cdot 10^{-9}$	8.22365	64	$3.32864 \cdot 10^{-9}$	15.8385
128	$2.34908 \cdot 10^{-10}$	8.10962	128	$2.08388 \cdot 10^{-10}$	15.9732
256	$2.91656 \cdot 10^{-11}$	8.05426	256	$1.30242 \cdot 10^{-11}$	16.0000
512	$3.63344 \cdot 10^{-12}$	8.02700	512	$8.14147 \cdot 10^{-13}$	15.9974

Unfortunately the approximation order is again 3. Analyzing carefully the distribution of the error over the splitting intervals, we realized that the greatest error is now not any more at the interval containing the inflection, but at the next one. Analyzing the behavior of the prototype support function $h = s^{3/2}$, we suggested an adapted subdivision scheme, where the first interval containing the inflection is not divided equally but in the ratio $(l : l^{5/4})$, where l is the length of the interval. This way, keeping the same number of segments, we obtain better error behavior with the approximation order 4.

5 Conclusion

We have analyzed the behavior of the support function in the neighborhood of a curve inflection. This analysis lead us to design an interpolation scheme which has the same

approximation order 4 both at inflection and non-inflection points. Using this scheme we are able to convert any curve into a curve with rational offsets with high precision. Such approximation may be particularly useful for CNC manufacturing. In the future we plan to investigate the inflections of higher order and to apply this approach to the approximation of algebraic curves with complicated singularities. We also plan to study the support function of surfaces at their parabolic lines.

References

[1] Aigner, M., Jüttler, B., Gonzalez-Vega, L. , Schicho, J.: Parametrizing surfaces with certain special support functions, including offsets of quadrics and rationally supported surfaces. Journal of Symbolic Computation 44, 180–191 (2009).

[2] Aigner, M., Gonzalez-Vega, L., Jüttler, B., Sampoli, M.L.: Computing isophotes on free-form surfaces based on support function approximation. In: Hancock, E., Martin, R., (eds.), The Mathematics of Surfaces (MoS XIII 2009), LNCS, vol. 5654, pp. 1–18. Springer (2009).

[3] Bastl, B., Lávička, M., Šír, Z.: G^2 Hermite Interpolation with Curves Represented by Multi-valued Trigonometric Support Functions. LNCS, vol. 6920, 142–156 (2012).

[4] Gravesen, J.: Surfaces parametrised by the normals. Computing 79, 175–183 (2007).

[5] Gravesen, J., Jütter, B., Šír, Z.: Approximating Offsets of Surfaces by using the Support Function Representation. In: Bonilla, L.L., Moscoso, M., Platero, G., Vega, J.M. (eds.): Progress in Industrial Mathematics at ECMI 2006, Mathematics in Industry 12, pp. 719-723. Springer Verlag (2007).

[6] Gravesen, J., Jüttler, B., Šír, Z.: On rationally supported surfaces. Comput. Aided Geom. Design 5 (4-5), 320–331 (2008).

[7] Gruber, P.M., Wills, J.M.: Handbook of convex geometry. North–Holland, Amsterdam (1993).

[8] Lávička, M., Bastl, B., Šír, Z.: Reparameterization of curves and surfaces with respect to convolutions. In: Dæhlen, M. et al. (eds.): MMCS 2008, LNCS, vol. 5862, pp. 285–298, Springer-Verlag Berlin, Heidelberg (2010).

[9] Sabin, M.: A Class of Surfaces Closed under Five Important Geometric Operations. Technical Report VTO/MS/207, British Aircraft Corporation (1974). Available at `http://www.damtp.cam.ac.uk/user/na/people/Malcolm/vtoms/vtos.html`

[10] Šír, Z., Bastl, B., Lávička, M.: Hermite interpolation by hypocycloids and epicycloids with rational offsets. Computer Aided Geometric Design 27, 405–417 (2010).

[11] Šír, Z., Gravesen, J., Jüttler, B.: Curves and surfaces represented by polynomial support functions. Theoretical Computer Science 392, 141–157 (2008).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The influence of distributed delays on Hes1 gene expression model

Marek Bodnar[1]

[1] *Institute of Applied Mathematics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland.*

emails: mbodnar@mimuw.edu.pl

Abstract

In the Hes1 gene expression system the protein (in fact dimmers of the protein) bounds to the promoter of its own DNA blocking transcription of its mRNA. As a result of such negative feedback loop an oscillatory behaviour is observed experimentally. The classical model that describes this system consists of two ordinary differential equations with discrete time delay in the term that reflects transcription. However, transcription take place in the nucleus while translation in the cytoplasm. This means that delay present in the system is in fact larger than transcription time. Moreover, it is somehow distributed around some mean value. During the pretension the model of the Hes1 gene expression system is presented. The similarities and differences between the model with discrete and distributed delay is discussed. It turns out that in the case of distributed delay the steady state is more stable than in the case of discrete delay.

Key words: distributed delays, gene expression model, stability analysis
MSC 2000: AMS codes (optional)

1 Introduction

There exist a number of genes that change their expression pattern in an oscillatory manner. In some cases these oscillations are stable and can be treated as molecular clocks as in circadian clock and the cell cycle (see e.g. [1, 7]). There are also reports on the oscillatory behaviour in the system with protein p53 after induction by DNA damage (see [8] for details). In 2002, Hirata *et al.* [9] observed oscillations in the Hes1 system. In 2003 Monk proposed a very simple model of this system with time delay, see [11]. Independently, Jensen *et al.* [10]

numerically studied the same model as proposed by Monk, and observed that sustained oscillations may be induced by time delay introduced to the system.

The mechanism of the operation of the Hes1 protein involves a negative feedback loop with one activation and one repression (see Fig. 1). The synthesis (transcription) of the mRNA of Hes1 activates the production (translation) of the Hes1 protein. On the another hand Hes1 represses the transcription of its own mRNA by bounding to DNA (see [9] for more details).

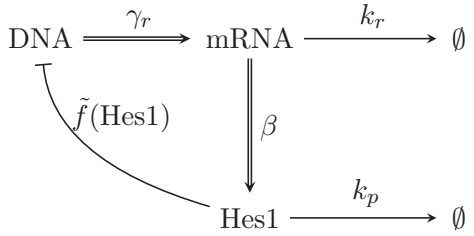


Figure 1: A sketch of negative feedback loop for the Hes1 system.

A classical model of this regulatory system proposed by Monk [11] reads

$$\begin{aligned} \dot{r}(\tilde{t}) &= \tilde{f}(p(\tilde{t}-\tau_r)) - k_r r(\tilde{t})\,, \\ \dot{p}(\tilde{t}) &= \gamma_p r(\tilde{t}-\tau_p) - k_p p(\tilde{t})\,, \end{aligned} \tag{1}$$

where p and r are the concentrations of Hes1 and its mRNA. It is worth pointing out that the probability of DNA being in an active state does not appear explicitly in the model (1), because it is assumed that this probability is approximately constant (for fixed concentration of the Hes1 protein) and therefore it is hidden in the function $\tilde{f}$. Throughout the paper we assume that the function $\tilde{f}$ is decreasing. The biological interpretation of this model is as follows:

- Hes1 protein is produced by its mRNA with intensity γ_p; we assume that Hes1 production is proportional to the concentration of its mRNA.

- mRNA and Hes1 are degraded proportionally to their concentrations, which corresponds to the terms $-k_r r$ and $-k_p p$; $1/k_r$ and $1/k_p$ are characteristic time for the degradation of mRNA and Hes1 protein, respectively — they can be also considered as mean life time these molecules.

- The term $\tilde{f}(p)$ describes a negative feedback loop. The hes1 dimmers bound to DNA blocking the production of his own mRNA. Thus, the intensity of production of hes1 mRNA decreases with increasing concentration of the Hes1 protein.

- We assume that time delay τ_r in the mRNA production term reflects the duration of molecular processes connected with transcription as well as the diffusion time of molecules between nucleus and cytoplasm, while τ_p reflects the duration of molecular processes connected with translation and also diffusion time.

In [10] a particular form of the function f where considered, namely

$$\tilde{f}(p) = \frac{\gamma_r k^h}{k^h + p^h} \tag{2}$$

where h is the Hill coefficient that takes into account the cooperative character of the binding process and k is a characteristic concentration for dissociation of Hes1 from DNA.

The model (1) was studied in several papers. In Bernard *et. al* [2] the system with delay in transcription process was analysed. In [5] the model with delay in both transcription and translation processes was studied. Theorem guaranteeing local stability of a unique steady state was proved as well as the existence of the Hopf bifurcation was shown. Moreover, the direction of the Hopf bifurcation was calculated. In [4] the global stability of the steady state was studied. On the other hand, Zeiser [13] studied the system assuming that the DNA promoter has three binding sites. Using a quasi-stationary approximation a particular form of the function $\tilde{f}$ was derived and it was approximated by a Hill function of the form (2). Sturrock *et al.* [12] considered the Hes1 system without time delay but taking into account diffusion of protein and mRNA between the cytoplasm and the nucleus.

In all papers mentioned above, a discrete delay was considered. Although time of transcription and translation can be considered as approximately constant, the diffusion time varies more. Thus, it is reasonable to assume that delay is distributed around some fixed value. In this paper we investigate the influence of delay distribution on the dynamic of model (1), in particular, on the stability of the steady state.

In order to simplify calculation we choose rescaling considered in [4]. The assumption that $\tilde{f}$ is decreasing implies that the equation

$$\tilde{f}(\xi) - \frac{k_p\, k_r}{\gamma_p}\xi = 0 \tag{3}$$

has a unique positive solution. Let denote it by $\bar{p}$. The system considered here reads

$$\begin{aligned} \dot{x}(t) &= f\left(\int_0^\infty \theta(s) y(t+s)\,\mathrm{d}s\right) - x(t)\,, \\ \dot{y}(t) &= x(t) - \mu y(t)\,, \end{aligned} \tag{4}$$

where θ is a probability distribution defined on $[0,+\infty)$ and

$$\begin{aligned} &f(\xi) = \frac{\gamma_p}{\bar{p}k_r^2}\,\tilde{f}(\bar{p}\xi) \ \text{ for all } \ \xi \in [0,+\infty)\,, \qquad x(t) = \frac{\gamma_p}{\bar{p}k_r} r(t), \\ &y(t) = \frac{1}{\bar{p}} p(t), \qquad \mu = \frac{k_p}{k_r}, \qquad t = k_r \tilde{t}. \end{aligned} \tag{5}$$

 ISBN: 978-84-608-6082-2

2 Model properties

In the case of discrete delays, that is when θ is a Dirac delta measure concentrated at the point τ, the system (4) is well studied in [2, 5, 4] and can be summarised as follows.

- Non-negative solutions to (4) globally exist and are bounded. Moreover, there exists a set that is invariant under the evolution of system (4).
- A point $(\mu, 1)$ is a unique non-negative steady state of system (4).
- If $\mu > |f'(1)|$, then the steady state $(\mu, 1)$ is locally asymptotically stable for all $\tau_1, \tau_2 \geq 0$.
- If $\mu < |f'(1)|$, then the steady state $(\mu, 1)$ is locally asymptotically stable for $\tau_1 + \tau_2 < \tau_{\rm cr}$, and it is unstable for $\tau_1 + \tau_2 > \tau_{\rm cr}$, where

$$\tau_{\rm cr} = \frac{\arccos\left(\frac{\mu-\omega_0^2}{f'(1)}\right)}{\omega_0}, \quad \omega_0 = \sqrt{\frac{1}{2}\left(-(1+\mu^2) + \sqrt{\left(1-\mu^2\right)^2 + 4\left(f'(1)\right)^2}\right)} \tag{6}$$

  At $\tau_{\rm cr}$ the Hopf bifurcation occurs.
- If $f'''(1) < -\alpha_{sc}\left(f''(1)\right)^2$, then the Hopf bifurcation is supercritical. The coefficient α_{sc} depends on $f'(1)$, and model parameters, see [4].
- For the Hill function f, that is if $\tilde{f}$ is given by (2), the existence of the Hopf bifurcation (for $\tau_1 + \tau_2$ sufficiently large) depends on the Hill coefficient h and the parameter $\kappa = kk_pk_r/(\gamma_p\gamma_r)$. If $\kappa \leq 1$ then there exists one critical value $h_{\rm cr,1}$ such that for $h < h_{\rm cr,1}$ the positive steady state of (1) is locally asymptotically stable for all delays while for $h > h_{\rm cr,1}$ the steady state looses its stability for large delays. If $1 < \kappa < \kappa_{\rm cr}$, then there exist two one critical values $1 < h_{\rm cr,1} < h_{\rm cr,2}$ such that for $h < h_{\rm cr,1}$ or $h > h_{\rm cr,2}$ the steady state is stable for all delays while for $h_{\rm cr,1} < h < h_{\rm cr,2}$ the steady state looses its stability for large delays. The formula for $\kappa_{\rm cr}$, that can be found in [3], is the following

$$\kappa_{\rm cr} = \frac{4\,\mathrm{e}^{-2}}{W_p(2\,\mathrm{e}^{-2})\Big(W_p(2\,\mathrm{e}^{-2}) + 2\Big)} \approx 1.12118,$$

  where the W_p is the Lambert W function (or Omega function). Here, we used uncalled coefficients. Because the scaling (5) depends on the function $\tilde{f}$ and thus on the Hill coefficient, similar formulas in expressed in terms of scaled coefficient μ and the function f would imply that the dependence on the Hill coefficient would be implicit.

Let us first formulate assumptions on the function f

(A1) $f\colon \mathbb{R}_{\geq} \to R_{\geq}$ is decreasing C^1 class function, that is $f'(x) < 0$ for all $x \in \mathbb{R}_{>}$, fulfilling $f(1) = \mu$.

and on the delay distribution θ

(A2) θ is a probabilistic measure defined on $\mathbb{R}_{\geq}$ with a finite expectation

$$\tau_{\text{av}} = \int_0^{+\infty} s\theta(s)\,\mathrm{d}s. \tag{7}$$

Now, for the case of distributed delay we may prove the solutions to (4) exist, are unique and defined for all $t \geq 0$ (in an appropriate chosen phase space of continuous functions). Moreover, if the initial functions are positive, solutions stay also positive.

Due to chosen scaling the steady state is the following $(\mu, 1)$, and the characteristic equation for system (4) around the steady state reads

$$W_\tau(\lambda) = \lambda\mathbb{I} - \det\begin{bmatrix} -1 & -d_1\hat{\theta}(\lambda) \\ 1 & -\mu \end{bmatrix} = \lambda^2 + (\mu+1)\lambda + \mu + d_1\hat{\theta}(\lambda), \tag{8}$$

where $\mathbb{I}$ is an identity matrix 2×2, $d_1 = -f'(1)$, and

$$\hat{\theta}(\lambda) = \int_0^{\infty} \theta(s,\tau)\,\mathrm{e}^{-\lambda s}\,\mathrm{d}s$$

is a Laplace transform of θ.

It turns out, that in the plane $(\mu, |f'(1)|)$ for which the steady state is stable independently on the magnitude of average delay is the smallest in the discrete delay case. In order to do that we consider a family of probabilistic measures indexed by some positive parameter τ. Precisely, we assume that

(A3) the family of measures $\left\{\theta(\cdot,\tau)\right\}_{\tau\in\mathbb{R}_{\geq}}$ is such that for any $\tau \in \mathbb{R}_{\geq}$ the measure $\theta(\cdot,\tau)$ fulfils (A2), the function

$$(\lambda,\tau) \mapsto \hat{\theta}(\lambda,\tau) = \int_0^{\infty} \theta(s,\tau)\,\mathrm{e}^{-\lambda s}\,\mathrm{d}s$$

is a continuous function in $\mathbb{C} \times \mathbb{R}^+$, and $\hat{\theta}(\lambda, 0) = 1$, for all $\lambda \in \mathbb{C}$.

Note that for $\tau = 0$ (8) reads

$$W_0(\lambda) = \lambda^2 + (\mu+1)\lambda + \mu + d_1$$

which implies that the steady state is locally asymptotically stable. For $\tau > 0$ by W_τ we denote the characteristic function (8) with $\theta(s) = \theta(s,\tau)$.

Looking for zeros of characteristic function the following theorems can be proved:

 ISBN: 978-84-608-6082-2

Theorem 2.1 *Let the function f fulfils condition* (A1) *and the arbitrary family of measures $\theta(\cdot,\tau)$ fulfils condition* (A3). *Moreover assume that the steady state $(\mu,1)$ of* (4) *is locally asymptotically stable for $\tau=0$ and an inequality $\mu>d_1$ holds. Then the steady state is stable for all $\tau\geq 0$.*

Theorem 2.2 *Let f fulfils condition* (A1) *and let θ fulfils* (A2). *If*

$$\tau_{av}<\frac{\mu+1}{d_1}. \tag{9}$$

where τ_{av} is defined by (7), *then the steady state $(\mu,1)$ of* (4) *is locally asymptotically stable.*

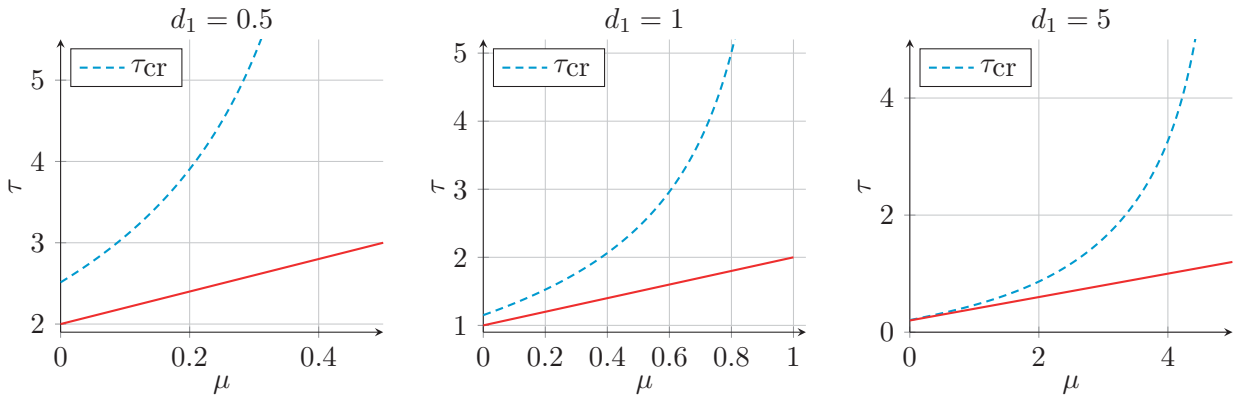


Figure 2: The comparison of the condition proved in Theorem 2.2 (the red solid line) and τ_{cr} for the case of discrete delays (the blue dashed line) for three different values of d_1.

For the specific delay distribution considered, that is Erlang distribution whic density is given by a formula

$$\theta(s)=\begin{cases}0 & s<\tau_m\\ \dfrac{a^m}{(m-1)!}(s-\tau_m)^{m-1}\,\mathrm{e}^{-a(s-\tau_m)} & s\geq\tau_m,\end{cases} \tag{10}$$

with $\tau_m\geq 0$, $a>0$ and $m\in\mathbb{N}$, we can prove an interesting result.

First, note that the Laplace transform of the Erlang distribution is

$$\hat{\theta}(\lambda)=\frac{a^m}{(a+\lambda)^m}\,\mathrm{e}^{-\lambda\tau_m}$$

and therefore in this case eigenvalues are solutions to the equation

$$\Big(\lambda^2+(\mu+1)\lambda+\mu\Big)\left(1+\frac{\lambda}{a}\right)^m+d_1\,\mathrm{e}^{-\lambda\tau_m}=0. \tag{11}$$

Now, using using Routh-Hurwitz criterion as well as Thorem 1 from [6] it can be proved:

Theorem 2.3 *Let the delay distribution θ be given by (10), assume that $\mu < d_1$ and that the steady state $(\mu, 1)$ of system (4) is locally stable for $\tau_m = 0$. Then there exists τ_c such that*

- *the steady state is stable for $0 \leq \tau_m < \tau_c$;*
- *the steady state is unstable for $\tau_m > \tau_c$;*
- *at $\tau_m = \tau_c$ the Hopf bifurcation occurs.*

Remark 2.4 *Note that Theorem 2.1 guarantees that for $\mu \geq d_1$ the steady state is always locally stable for all values of delay.*

The above result is similar in some sense to the discrete delay case. However, for the case of non-shifted exponential distribution, that is $m = 1$ and $\tau_m = 0$ we may formulate

Theorem 2.5 *Let the delay distribution θ be given by (10) with $m = 1$ and $\tau_m = 0$. If*

(i) *$d_1 < (\mu_1 + 1)\left(1 + \sqrt{\mu}\right)^2$ then the steady state $(\mu, 1)$ of system (4) is stable for all $\tau_{av} = 1/a$;*

(ii) *$d_1 > (\mu_1 + 1)\left(1 + \sqrt{\mu}\right)^2$ then there exists $0 < \tau_1 < \tau_2$ such that the steady state $(\mu, 1)$ of system (4)*

   (a) is stable for $1/a = \tau_{av} < \tau_1$ and $1/a = \tau_{av} > \tau_2$;
   (b) is unstable for $\tau_1 < \tau_{av} < \tau_2$,

   where

$$\tau_1 = \frac{2(\mu+1)}{d_1 - (\mu+1)^2 + \sqrt{\Delta_{RH}}}, \quad \tau_2 = \frac{2(\mu+1)}{d_1 - (\mu+1)^2 - \sqrt{\Delta_{RH}}} \tag{12}$$

   with

$$\Delta_{RH} = d_1^2 - 2(1+\mu)^2 d_1 + (1+\mu)^2(1-\mu)^2. \tag{13}$$

   The Hopf bifurcation occurs at τ_1 and τ_1.

For the case $m = 2$ the situation is more complex because the characteristic function is a polynomial of degree 4.

Proposition 2.6 *Let the delay distribution θ be given by (10) with $m = 2$ and $\tau_m = 0$. The following statements are true*

(i) *if $\mu < d_1 < 2\mu$ the steady state $(\mu, 1)$ of system (4) is stable for all $\tau = 2/a$;*

(ii) *if τ_{av} is large enough, the steady state of system (4) is stable.*

 ISBN: 978-84-608-6082-2

(iii) *if* $2\mu < d_1$ *there can exist at most two critical values* $\tau_1 < \tau_2$ *such that the steady state of system* (4) *is stable for* $\tau_{av} < \tau_1$ *and* $\tau_{av} > \tau_2$ *and it is unstable for* $\tau_1 < \tau_{av} < \tau_2$.

We want to emphasize that this situation differs from the discrete delay case as the steady state is stable for both small and large delays and it is unstable for moderate ones. The result of these theorems are illustrated at Fig. 3

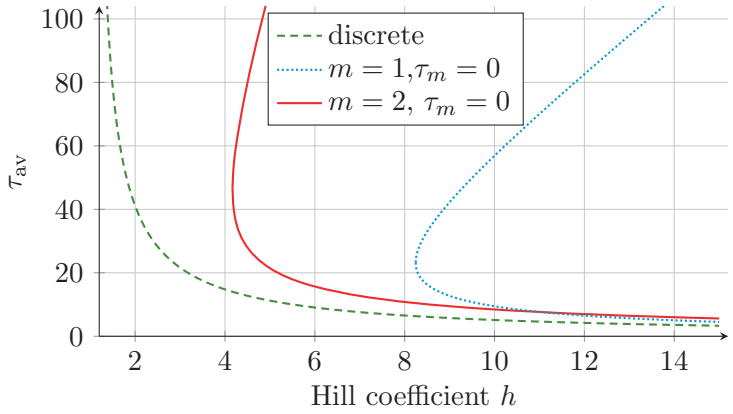


Figure 3: The dependence of the critical average delay value on the Hill coefficient. Time delay value is given before rescaling, in minutes. All parameter are as proposed by Monk [11]. The lines indicate critical average delay for different delay distributions: the dotted blue line for Erlang distribution with $m = 1$; the solid red line for Erlang distribution with $m = 2$; the dashed green line for discrete delay. The stability region is left to the curves.

3 Conclusions

The basic mathematical properties of the model with distributed delay are the same as in the case of discrete delay. For non-negative, continuous initial functions, solutions to (4) exits, are unique, non-negative and are defined for all $t \geq 0$. However, the stability of the steady states changes. The results presented in this paper indicated the smallest stability region corresponds to the discrete delay case. This was precisely stated (under some assumptions) in Theorem 2.1. Moreover, numerical simulations suggest that stability region for the distributed delay decreases with the decrease of the variance of the distribution.

On the other hand, for the exponential delay distribution (ie. the Erlang delay distribution with the shape parameter equal to 1) and Erlang delay distribution with the shape parameter equal to 2, the steady state is stable for both small and large value of average delay and it can be unstable for moderate values of average delays. This situation is differ-

ent from the discrete delay case when if the steady state looses its stability then it remains unstable for large value of delay.

Acknowledgements

This work has been supported by National Science Centre, Poland under the project no. 2015/17/B/ST1/00693.

References

[1] M. P. Antoch, E.-J. Song, A.-M. Chang, M. H. Vitaterna, Y. Zhao, L. D. Wilsbacher, A. M. Sangoram, D. P. King, L. H. Pinto, and J. S. Takahashi. Functional identification of the mouse circadian clock gene by transgenic BAC rescue. *Cell*, 89(5):655–667, 1997.

[2] S. Bernard, B. Cajavec, L. Pujo-Menjouet, M. C. Mackey, and H. Herzel. Modelling transcriptional feedback loops: the role of Gro/TLE1 in Hes1 oscillations. *Phil. Trans. R. Soc. A*, 364:1155–1170, 2006.

[3] M. Bodnar. Modele reakcji biochemicznych z opóźnionym argumentem: nieujemność rozwiązań i stabilność oscylacji. In A. Bartłomiejczyk, editor, *Metody matematyczne w zastosowaniach, tom 2*, chapter 1, pages 1–20. Centrum Zastosowań Matematyki, Politechnika Gdańska, 2014. in polish.

[4] M. Bodnar. General model of a cascade of reactions with time delays: global stability analysis. *J. Diff. Eqs.*, 259(2):777–795, 2015. arXiv:1403.5435.

[5] M. Bodnar and A. Bartłomiejczyk. Stability of delay induced oscillations in gene expression of Hes1 protein model. *Non. Anal. - Real.*, 13(5):2228–2239, 2012.

[6] K. L. Cooke and P. van den Driessche. On zeroes of some transcendental equations. *Funkcj. Ekvacioj*, 29:77–90, 1986.

[7] G. Fu, Z. Wang, J. Li, and R. Wu. A mathematical framework for functional mapping of complex phenotypes usingdelay differential equations. *J. Theor. Biol.*, 289:206–216, 2011.

[8] Y. Haupt, R. Maya, A. Kazaz, and M. Oren. Mdm2 promotes the rapid degradation of p53. *Nature*, 387(6630):296–299, 1997.

[9] H. Hirata, S. Yoshiura, T. Ohtsuka, Y. Bessho, T. Harada, K. Yoshikawa, and R. Kageyama. Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science*, 298:840–843, 2002.

 ISBN: 978-84-608-6082-2

[10] M. Jensen, K. Sneppen, and G. Tiana. Sustained oscillations and time delays in gene expression of protein Hes1. *FEBS Lett.*, 541:176–177, 2003.

[11] N. A. Monk. Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptional time delays. *Curr. Biol.*, 13:1409–1413, 2003.

[12] M. Sturrock, A. J. Terry, D. P. Xirodimas, A. M. Thompson, and M. A. Chaplain. Spatio-temporal modelling of the Hes1 and p53-Mdm2 intracellular signalling pathways. *J. Theo. Biol.*, 273(1):15–31, 2011.

[13] S. Zeiser, H. V. Liebscher, H. Tiedemann, I. Rubio-Aliaga, G. K. Przemeck, M. H. de Angelis, and G. Winkler. Number of active transcription factor binding sites is essential for the Hes7 oscillator. *Theor. Biol. Med. Modelling*, 3:11, 2006.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A tradeoff between explicit and implicit schemes to solve differential equations on GPUs

Marcelo Bondarenco[1], Pablo Gamazo[1] and Pablo Ezzatti[2]

[1] *Salto, Departamento del Agua, Cenur Noroeste, Universidad de la República, Uruguay*

[2] *Instituto de la Computación, Universidad de la República, Uruguay*

emails: {mbondarenco,mpgamazo}@unorte.edu.uy, , pezzatti@fing.edu.uy

Abstract

This paper addresses the resolution of the transport (advection–diffusion) equation in 3D using implicit and explicit schemes. Each scheme is implemented in two parallel codes (C+CUDA and C+OpenMP), and in sequential code (in C). The resolution of this equation using GPUs has been addressed by several authors, but to our knowledge, the comparison of precision and performance of implicit and explicit schemes on GPUs has not been studied before. The precision of each implementation is obtained by comparing results with an analytic solution.

Key words: GPU, differential equation, finite differences

1 Introduction

The solution of partial differential equations is required in several scientific areas such as structural mechanics, fluid dynamics and thermodynamics. In these applications, the use of numerical models is mandatory to address their solution.

The increase on the processing speed and the memory size of computers, accompanied by the development of modeling code, has allowed researchers to work with more complex physical models, even in desktop computers. However, in the last years the speed of computer processors has not increased much due to the physical limits of semiconductors. In the other hand, the graphics processors (GPUs) arose as an alternative resource on the high performance computing (HPC) field. Specifically, this kind of massively parallel processors are cheaper when compared with other HPC options (e.g. clusters), and offer an impressive computational power. Thus, nowadays the GPUs are widely used in scientific environments.

 ISBN: 978-84-608-6082-2

This new paradigm of HPC motivates the study of the relation between different numerical methods and the acceleration achieved by them on the GPU with respect to multi-core parallelism. Specifically, our focus is on the use of Finite Differences (FDM) technique to solve a system of partial differential equations, or ordinary differential equations, with boundary and initial conditions. It should be noted that FDM is one of the most suitable and widespread method to solve this kind of system equations. We can work with FDM using explicit or implicit schemes. Implicit schemes are typically more expensive than explicit methods but the latter achieve better levels of accuracy.

Concretely, in this paper we study the solution of the transport (advection–diffusion) equation as a workhorse. The equation is addressed in 3D using both implicit and explicit schemes. For each of them, two parallel codes (C+CUDA and C+OpenMP) and a sequential code (C) were implemented.

The rest of the paper is structured as follows. In Section 2 we review the FDM, its application to the transport equation and summarize some related works. Section 3 describes the most important details about our parallel implementations of the solvers. After that, in Section 4, the experimental evaluation of the different methods is presented, and finally, Section 5 offers a few concluding remarks.

2 Transport equation

As we stated previously, we use the transport equation as a workhorse. The transport phenomena are the processes in which there is an exchange of mass, energy or linear momentum. From a mathematical point of view, they are linear systems of partial differential equations with parabolic and hyperbolic terms as we can see in Equation 1, where C is a concentration [mass/meters^3], diffusion coefficient [meters^2/time] and v is a speed [meters/time].

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\frac{\partial (D \cdot C)}{\partial x} - \frac{\partial (v \cdot C)}{\partial x} \tag{1}$$

In this work we are focused on the 3D variant of the equation, and for the resolution we choose the finite differences technique as a discretization procedure. In particular, we evaluate both traditional schemes, i.e. explict and implict schemes.

Explicit schemes:

In this approach the values for each point i in time $t + \Delta t$ are computed with the values on different points, e.g. the i and some neighborhood, in time t. The use of different neighbor definition implies a different lattice of computation. The method offers a simple computation strategy, but restricts the selection of the Δt. In particular, for stability reasons in some cases the method requires the use of extremely small values of Δt. One discretization of the Equation 1, based on an explicit scheme is:

$$\frac{f^{t+\Delta t}(x) - f^t(x)}{\Delta t} = D \cdot \left(\frac{f^t(x+h) - 2f^t(x) + f^t(x-h)}{h^2}\right) - v \cdot \left(\frac{f^t(x+h) - f^t(x)}{h}\right). \tag{2}$$

And, in this case, the stabilization conditions is: $\frac{D \cdot \Delta t}{h^2} \leq \frac{1}{2}$.

Implicit schemes:

One strategy to overcome stability issues arising in the finite differences context is the use of implict schemes. This scheme employs values in time $t + \Delta t$ to compute the values for a point i in time $t + \Delta t$. Therefore, the use of this approach implies the solution of one linear equation system solver on each time step. One formulation for this approach is:

$$\frac{f^{t+\Delta t}(x) - f^t(x)}{\Delta t} = D \cdot (\frac{f^{t+\Delta t}(x+h) - 2f^{t+\Delta t}(x) + f^{t+\Delta t}(x-h)}{h^2}) - u \cdot (\frac{f^{t+\Delta t}(x+h) - f^{t+\Delta t}(x)}{h}). \quad (3)$$

2.1 Related work

The use of the computational power of GPUs to accelerate the solution of the transport equation has been addressed for several efforts in last years. Some highlighted examples are described next. First, Goncalves [1] perform an evaluation of several methods to solve the linear systems from 1-D variant of the problem. Later, in Cotronis et al. [2], the authors applied the successive over-relaxation (SOR) method, as a linear system solver, for the implicit variant of finite differences. Finally, we highlight the work by Molnár et al. [5] where the solution of the reaction–diffusion equation is addresed.

3 Implementation

In this section we offer a brief abou tonation of our parallel implementations for the transport equation solvers, following both the explicit and implicit schemes.

3.1 Explicit scheme

This scheme presents few data dependencies, and the parallelization is only limited by the number of processing units (and the relation between computing and memory access, i.e. in general these are memory bound scenarios). To address this technique we divide the processing in 2 stages, first we compute the interior points and, later, the border points are calculated. More in detail we have 2 versions:

- Expl$_{cpu}$: We divide the grid to map different zones with CPU threads using the OpenMP API.
- Expl$_{gpu}$: The GPU implementation is similar to the previous one, but we use CUDA as a parallel tool.

3.2 Implicit scheme

The most important stage, from the computational point of view, of this strategy is the solution of a linear equation system for each time step. To tackle the linear systems we

employ the SIP [4]. SIP is an iterative method for the resolution of band systems ($Ax = b$, where A is a penta-diagonal or hepta-diagonal matrix), that are usually derived from the resolution of differential equations schemes with regular grids. SIP method is described next. Initially, a special incomplete LU factorization ($\hat{L}\hat{U} = incLU(A)$), guided by an α parameter to improve the approximation is computed, where $\hat{L}$ and $\hat{U}$ are good approximations of L and U matrices without fill-in. After that, an iterative procedure that refines an initial solution is performed until the residual is small enough. A good value for α parameter is 1.8 (more details can be found in [4]). We parallelize the SIP for both platforms:

- IMPL$_{cpu}$: Following the ideas presented in Darseno et al. [3], we perform the processing of several points of the grid concurrently by grouping by hyper-planes and leveraging the OpenMP API.
- IMPL$_{gpu}$: The GPU implementation is similar to the previous one, but we use CUDA as a parallel tool.

4 Experimental evaluation

The target CPU-GPU server containes an Intel Xeon(R) CPU E5-2620v2 (6 cores at 2.1 GHz, Sandy-Bridge architecture) with 128 GB of DDR3 memory, connected to a NVIDIA K40 GPU (2880 CUDA cores at 745 MHz, Kepler architecture, 12 GB of DDR5 memory) via a PCI-e bus. The operating system was Centos v6.5 Linux. The codes were compiled using gcc v.4.4.7 and the NVIDIA CUDA compiler v.6.5 (both with the -O3 flag enabled).

As a case study, we solve the transport equation on a rectangular prism with $258 \times 130 \times 130$ dimension. This is a commonly used test case in the fluid dynamics field. All experiment are performed using double precision arithmetics and, in the GPU cases, the transfer times are always included.

The first study evaluate the runtime performance of our four parallel version and the comparision with the sequential CPU variants. In this line, Table 1 summarizes the runtime requiered to tackle our test case. Considering the hardware platform for the parallel CPU version we employ 6 threads (the number of fisical cores). On the other hand, the residual error results, calculated as the norm 2 of the differences between the numerical and the analitic solution, present singnificant differences. In concrete, the explicit paradigm reachs an error of 3.26E-03 while the implicit paradigm yields an error value of 3.35E-04.

The obtained results confirm that the explicit scheme is cheaper than the implicit counterpart, but the results from the implicit approach offer a better accuracy. In other line, and based on the parallelism application, the parallel performance of the CPU for both methods is poor, which can be explained by the high number of memory access (both method are memory bound). Nevertheless, the use of the GPU for explicit schemes shows better resulta than its use for implicit schemes (note that in the first case the speedup is near

	CPU seq. time	CPU par. time	GPU time
Explicit	8.00E+01	3.86E+01	1.16E+01
Implicit	4.76E+02	3.68E+02	2.53E+02

Table 1: Runtimes (in seconds) for different method to tackle the transport equation.

to 7×, while in the implicit case is only up to 2×). This result is important because, with these values of acceleration, the accurancy problems of explict schemes can be compensated by using a finer grid for the computations.

5 Concluding Remarks

This work presented a study about the acceleration of the resolution of differential equation with GPUs. Specifically, we used the solution of the transport (advection–difusion) equation in 3D as a workhorse to compare the use of implicit and explicit schemes. Each scheme was implemented in two parallel codes (C+CUDA and C+OpenMP), and in one sequential code (in C), and we compared the precision and performance of the different approaches.

The experimental evaluation conducted showed that implicit methods offer better accurracy resultsbut the use of GPUs allows higher values of acceleration in the explict case.

As part of future work, we plan to study the behavior of both schemes in other test cases. Also, we intend to study other paradigms, e.g. predictor–corrector schemes.

References

[1] Gil Gonçalves Brandao. *Solution of the Transport Equation using Graphical Processing Units*. PhD thesis, MSc thesis IST, 2009.

[2] Yiannis Cotronis, Elias Konstantinidis, and Nikolaos M Missirlis. A gpu implementation for solving the convection diffusion equation using the local modified sor method. In *Numerical Computations with GPUs*, pages 207–221. Springer, 2014.

[3] F. Deserno, G. Hager, F. Brechtefeld, and G. Wellein. Basic optimization strategies for cfd-codes. *Technical report, Regionales Rechenzentrum Erlangen*, 2002.

[4] J.H. Ferziger and M. Perić. *Computational methods for fluid dynamics*. Numerical methods: Research and development. Springer-Verlag, 2002.

[5] Ferenc Molnar, Ferenc Izsak, Robert Meszaros, and Istvan Lagzi. Simulation of reaction–diffusion processes in three dimensions using cuda. *Chemometrics and Intelligent Laboratory Systems*, 108(1):76–85, 2011.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Auto-Tuning TRSM with an Asynchronous Task Assignment Model on Multicore, GPU and Coprocessor Systems

Murilo Boratto[1], Pedro Alonso[2], Pau San Juan[2] and Domingo Giménez[3]

[1] *Núcleo de Arquitetura de Computadores e Sistemas Operacionais,*
Universidade do Estado da Bahia

[2] *Departament de Sistemes Informàtics i Computació,*
Universitat Politècnica de València

[3] *Departamento de Sistemas Informáticos,*
Universidad de Murcia

emails: muriloboratto@uneb.br, palonso@upv.es, p.sanjuan@upv.es, domingo@um.es

Abstract

The development of high performance dense linear algebra depends on highly optimized BLAS and LAPACK, one of whose main kernels solves triangular linear systems (TRSM). Hence, the increasing necessity of computational power today justifies the large advances in searching for techniques that decrease the time to answer usual computational problems. To take advantage of the new computational environment, our current research shows to incorporate algorithms with high parallelism characteristic, to efficiently account for the multithreading and multiprocessor architectures available. In this paper, we propose an automatic tuning methodology to highly and easily exploit the multicore, GPU and coprocessor systems. So, we present an optimization of algorithm for solving triangular systems, employing block decomposition with an asynchronous task assignment model.

Key words: Automatic Tuning, Performance, Multicore, GPU, Coprocessor

1 Introduction

As processors become faster, the integrated circuit manufacturing companies clearly need to produce higher performance processors. This rise in performance is basically achieved

by decreasing component sizes and increasing processor clock speed. Soon physical components will be so small there will be no further space for improvement. Manufacturers have therefore been obliged to create alternatives to single processors, incorporating multicore and manycores structures. The processor presents a performance gain based on the number of available cores [4], meaning that in a future personal computers, servers, embedded systems, etc, will be rigged with high capacity.

It is up to software developers to create superscalar programs, since it is highly probable that a program written today will need to be totally modified tomorrow to adapt to machines with more processors [1]. To meet this need, manipulation patterns were created which allow modifications in existing codes by incorporating directives. This means users do not need to radically update their applications to enjoy the benefits of multiprocessing environments.

Further algorithmic advances are needed to use all the cores efficiently, especially since the number of cores in a socket is likely to double every two years. The performance of threaded algorithms seems to be related to fine granularity and asynchronicity. Fine granularity requires splitting an operation into tasks that operate on small portions of data to reduce bus traffic and improve data locality [8]. Asynchronicity avoids the presence of synchronization points that seriously affect the efficiency of the concurrent execution of the algorithm.

We choose triangular linear systems solver (TRSM) as representative level BLAS [3] routine to implement the asynchronous task assignment model. The TRSM routine solves the linear equation $AX = B$, where A is an upper or lower triangular matrix and B is a known called right-hand side matrix. Its implementation involves a block decomposition algorithm in parallel followed by a series of operations to get the time distribution of the multicore, GPU and coprocessor systems, where the solution can be expressed as two parts: solve a small triangular system with TRSM and block matrix multiply with GEMM.

In this paper we present an optimization of the algorithm for solving triangular systems (TRSM). We apply a technique for the automatic tuning of workload distribution among asynchronous tasks on shared memory architectures. The main contributions of this paper and possible extensions of the work are outlined. The aim is to improve the TRSM routine and also to use it as a testbed for other routines which can benefit from the same automatic tuning methodology that highly and easily exploits a multicore, GPU and coprocessor systems. The rest of the paper is organized as follows: Section 2 shows the related work to automatic tuning. Section 3 presents experimental results. The conclusions section closes the paper.

2 Related work

There exist important automatic tuning systems that attempt to automatically adapt the software to tune to the conditions of the execution platform. These include, e.g. the

FFTW package developed for the computation of discrete Fourier transform [7], ATLAS for BLAS [16], a library of linear algebra routines for sparse matrices [9], etc. The main goal of any automatic tuning system is to minimize the execution time of the routine to tune, keeping in turn the installation time below a reasonable threshold. In addition, the existence of automatically tuned software makes easy the efficient utilization of the library routines by non-expert users.

The approach chosen, e.g. by FAST [5], consists of a large benchmark followed by a polynomial regression to find the optimal parameters for different routines. Polynomial regression is used in [15] to decide which is the most appropriate version among the different variants of a routine. The authors of the same work also introduce a black-box running method to reduce the enormous implementation space. In the approach followed by FIBER [10] the execution time of a routine is approximated by fixing one parameter and varying the other one. In this case, a set of polynomial functions of grades $1-5$ is generated and the best one of them all is selected. The values provided by these functions for different problem sizes are then used to generate another function where the second parameter is fixed now and the first one is varied.

The solution of triangular systems (TRSM) were optimized in [6] by the use of the numerical library, where matrices were converted to floating point representations. At the present time, the current hardware trends have inevitably brought the need for updates on existing legacy software packages, such as BLAS. This is reflected, for instance, in the *Parallel Linear Algebra Software for Multicore Architectures* (PLASMA) project [14], and the *Matrix Algebra on GPU and Multicore Architectures* (MAGMA) project [12], which is a recent effort on developing a LAPACK [2] version for multicore and heterogeneous/hybrid architectures containing hardware accelerators like GPUs. The goal of this work falls within the context of all these above mentioned packages which try to build the best routine through the selection of some critical parameters at installation time.

3 Performance results

This section presents the experiments of the application to the linear system routine of the automatic tuning methodology. Experimental results are shown and commented, with detailed explanation and useful insights, with the following execution environment:

> **[Platform]** Execution environment with 2 identical GPUs. Comprises 2 Intel Xeon at 3 GHz and 96 GB DDR3 main memory. Each one is a hexacore processor (Hyper-Threading is set) with 24 MB of cache memory. It contains two GPUs NVIDIA Tesla C2050 with 14 stream multiprocessors (SM) and 32 stream processors (SP) each (448 cores in total). A Many Integrated Core (MIC) Intel Xeon Phi, with 57 cores at 1.1 GHz based on Pentium (x86), each core supporting 4 hardware threads, with a bidirectional ring bus and up to 6 GBytes GDDR5.

Table 1: Execution time obtained at installation time, with different parameters values (in seconds).

Platform	$n = 10240$		$n = 11200$		$n = 12800$		$n = 14400$	
c	w	$t(n, c, w)$	w	$t(n, c, w)$	w	$t(n, c, w)$	w	$t(n, c, w)$
1	16	97.86	16	125.64	16	193.59	32	270.78
2	16	52.90	16	70.06	32	105.80	32	173.19
4	16	28.79	32	39.23	32	59.22	64	85.46
6	32	17.46	32	22.87	32	33.60	64	48.42
8	32	14.25	32	19.09	32	29.71	64	41.99
10	64	13.24	64	17.85	64	27.87	64	38.93
12	64	12.28	64	14.74	64	23.24	64	32.16
14	64	13.76	64	16.50	64	26.45	64	35.65
16	64	15.04	64	17.73	64	28.04	64	33.86

We have implemented a parallel algorithm using MAGMA and OpenMP [13] which is based on the block decomposition and n the experiments, the MKL library [11] is used for BLAS routines in the system, and an optimum split of the matrices would keep the time consumed by cores balanced. The application is executed with an `icc` compiler version 12.1.3. Many parameter values were used at installation time to estimate the AP values using a number of cores (c), the range available being from 1 to 16, with an increase value of 2 in Platform, to obtain the number that minimizes the time. Then, we checked that the block size (w) from 1 to $1,024$, with an increase value of 2. The input sizes of the problem (n) for the experiments from $10,240$ to $32,000$, with an increase value of $2,048$. Table 1 shows the parameters used at installation time to estimate the AP values for the environment. The installation time spent on system was around 7 minutes. We did experiments with different combinations of n, c and w, considering the small size problem to obtain the model on a given platform. There are two important observations: (1) the c values depend on the problem size in the system under test, and (2) for each problem size and for different values of w we obtain a different optimum value for c on disparity execution environments. Being aware of this variability is fundamental to make good decisions in the later selection of the optimum AP parameters.

In Table 1, obtained values are shown for the experimental platform. After varying the amount of number of cores in experiments, we observed that performance improved by using the optimum block size established in automatic tuning installation (around w=64 in Platform). The optimization in the execution time is justified by the cache-level exploring nature of hybrid architectures which perfectly adjusts to the optimum block size found. The table shows the execution platform in installation stage with a potential increase due to the number of cores of the tested node. Looking through the table for optimum experimented block size, we can see that task calculation time decreases with the number of cores used.

This reduction in time occurs until we reach around 12 threads when the number of cores is 12 close to this range available in Platform. Moreover, asynchrony provides algorithms with a good scalability, having in the management of performed tasks, a workload balance between thread execution.

4 Conclusions and future directions

The algorithm encounters its foundation in the idea of unwrapping all of the potential implemented into BLAS routines teamed with the developed an asynchronous task assignment model. It is applied to obtain balanced distributions of the work to execute TRSM routine. Experiments are conducted for a number of problem sizes and varying the size of the matrices partition when the routine is being installed in the system. It is necessary to guide the search for the preferred partition to reduce the installation time, especially if the number of cores increases.

The experimental results obtained in this paper indicate that the proposed model is efficient, based on task classification, response time and workload distribution. The model demonstrates the amount of computational power of hybrid architecture, which is achieved combining the intrinsic parallelism of the algorithm, adaptations and software tools that we used during execution. As always, there is still room for improvement in this field, like for example, improving the assignment stage with a faster method. Another suggestion could perhaps be the inclusion of alternative programming models that would make transformation of legacy code for multicore, GPU and coprocessor systems possible.

Finally, more experiments in more systems with larger numbers of cores of different architectures and with other linear algebra routines are needed, and the adaptation of a technique based on theoretical models should be analyzed. More complex systems should be considered, for example multicore+multi-GPU+multi-MIC and heterogeneous clusters with heterogeneous nodes.

References

[1] S. Akhter and J. Roberts. *Multi-core programming*, volume 33. Intel press Hillsboro, 2006.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK users guide*. SIAM, Philadelphia, 2 edition, 2013.

[3] BLAS. Basic Linear Algebra Subprograms. Available in: http://www.netlib.org/blas/forum/, January 2016.

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

[5] E. Caron and F. Uter. Parallel extension of a dynamic performance forecasting tool. *Sci Ann Cuza Uni*, 11:80–93, 2002.

[6] J. Dumas, P. Giorgi, and C. Pernet. FFPACK: Finite field linear algebra package. In *IISAC04 International Symposium on Symbolic and Algebraic Computation*, pages 119–126, New York, NY, USA, 2004. ACM.

[7] M. Frigo and S. Johnson. FFTW: an adaptive software architecture for the FFT. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, volume 3, pages 1381–1384, May 1998.

[8] R. Karmani, P. Madhusudan, and B. Moore. Thread contracts for safe parallelism. In *Proceedings of the 16th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP 2011, San Antonio, TX, USA, February 12-16, 2011*, pages 125–134, 2011.

[9] T. Katagiri, K. Kise, and H. Honda. RAO-SS: a prototype of run-time auto-tuning facility for sparse direct solvers. Technical report, June 2005.

[10] T. Katagiri, K. Kise, H. Honda, and T. Yuba. Effect of Auto-Tuning with user's knowledge for numerical software. In *Conf. Computing Frontiers*, pages 12–25, 2004.

[11] MKL. Intel Math Kernel Library. Available in: http://developer.intel.com/software/products/mkl/, January 2016.

[12] R. Nath, S. Tomov, and J. Dongarra. An Improved MAGMA GEMM for FERMI Graphics Processing Units. *International Journal of High Performance Computing Applications*, 24(4):511–515, November 2010.

[13] OpenMP. Open Multi-Processing. Available in: http://www.openmp.org, January 2016.

[14] PLASMA. Parallel linear algebra software for multicore architectures. Available in: http://www.netlib.org/plasma/, June 2015.

[15] R. Vuduc, J. Demmel, and J. Bilmes. Statistical models for empirical search-based performance tuning. *J High Perform Comput Appl*, 18:65–94, February 2004.

[16] R. Clint Whaley, A. Petitet, and Jack J. Dongarra. Automated empirical optimizations of software and the ATLAS project. *Parallel Comput*, 27:21–37, 2001.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Linear and Cyclic Codes over direct product of Finite Chain Rings

Joaquim Borges[1], Cristina Fernández-Córdoba[1] and Roger Ten-Valls[1]

[1] *Department of Information and Communications Engineering,*
Universitat Autònoma de Barcelona

emails: joaquim.borges@uab.cat, cristina.fernandez@uab.cat, roger.ten@uab.cat

Abstract

We introduce a new type of linear and cyclic codes. These codes are defined over a direct product of two finite chain rings. The definition of these codes as certain submodules of the direct product of copies of these rings is given and the cyclic property is defined. Cyclic codes can be seen as submodules of the direct product of polynomial rings. Generator matrices for linear codes and generator polynomials for cyclic codes are determined.

Key words: Codes over rings, linear codes, cyclic codes, finite chain rings
MSC 2000: 94B60, 94B25

1 Introduction

Linear codes are a special family of codes with rich mathematical structure. One of the most studied class of linear codes is the class of linear cyclic codes. The algebraic structure of cyclic codes makes easier their implementation. For this reason many practically important codes are cyclic.

The study of codes over rings has been growing since it was proven in [8] that certain notorious non-linear binary codes can be seen as binary images under the Gray map of linear codes over $\mathbb{Z}_4$. In particular, the family of codes over chain rings has received much attention because it includes some good codes (e.g. [6], [9]).

In recent times, linear codes with sets of coordinates over different rings are studied (e.g. $\mathbb{Z}_2^\alpha \times \mathbb{Z}_4^\beta$ in [3], $\mathbb{Z}_{p^r}^\alpha \times \mathbb{Z}_{p^s}^\beta$ in [2]). Also, linear cyclic codes over these kind of structures are studied, see [1], [4], [5] and [7].

In this paper we present the structure of linear and cyclic codes over direct product of finite chain rings, $\mathcal{R}_1$ and $\mathcal{R}_2$. Linear codes can be seen as certain submodules of $\mathcal{R}_1^{\alpha} \times \mathcal{R}_2^{\beta}$, and cyclic codes as submodules of $\mathcal{R}_{\alpha,\beta} = \frac{\mathcal{R}_1[x]}{\langle x^{\alpha}-1\rangle} \times \frac{\mathcal{R}_2[x]}{\langle x^{\beta}-1\rangle}$. We determine the generator matrix in standard form and the generator polynomials in the cyclic case. Finally, we present examples to illustrate some particular cases.

2 Review of cyclic codes over finite chain rings

Let $\mathcal{R}$ be a finite chain ring with maximal ideal $\langle\gamma\rangle$ and let e be the nilpotency of γ. It is well-know that there exist a prime p and a positive integer m such that $|\mathcal{R}/\langle\gamma\rangle| = q = p^m$ and $|\mathcal{R}| = q^e = p^{me}$.

Let C be a cyclic code of length n over $\mathcal{R}$. It is known that we can identify C as an ideal of $\mathcal{R}[x]/(x^n-1)$. We assume that n is a positive integer such that it is coprime with p. Therefore, the polynomial x^n-1 has a unique decomposition as a product of basic irreducible polynomials that are pairwise coprime over $\mathcal{R}[x]$.

Theorem 2.1 ([6, Theorem 3.5])**.** *Let C be a cyclic code of length n over a finite chain ring $\mathcal{R}$, which has maximal ideal $\langle\gamma\rangle$ and e is the nilpotency of γ. Then, there exist polynomials $g_0, g_1, \ldots, g_{e-1}$ in $\mathcal{R}[x]$ such that $C = \langle g_0, \gamma g_1, \ldots, \gamma^{e-1} g_{e-1}\rangle$ and $g_{e-1} \mid g_{e-2} \mid \cdots \mid g_1 \mid g_0 \mid (x^n-1)$.*

Let $C = \langle g_0, \gamma g_1, \ldots, \gamma^{e-1} g_{e-1}\rangle$ be a cyclic code of length n and let $g = g_0 + \gamma g_1 + \cdots + \gamma^{e-1} g_{e-1}$. Since g_0 is a factor of x^n-1 and for $i = 1 \ldots e-1$ the polynomial g_i is a factor of g_{i-1}, we will denote $\hat{g}_0 = \frac{x^n-1}{g_0}$ and $\hat{g}_i = \frac{g_{i-1}}{g_i}$ for $i = 1 \ldots e-1$. Define $G = \prod_{i=0}^{e-1} \hat{g}_i$, then it is clear that $Gg = \left(\prod_{i=0}^{e-1} \hat{g}_i\right) g = 0$ over $\mathcal{R}[x]/(x^n-1)$.

Lemma 2.2. *Let C be a cyclic code of length n over a finite chain ring $\mathcal{R}$, which has maximal ideal $\langle\gamma\rangle$ and e is the nilpotency of γ. Let $g_0, g_1, \ldots, g_{e-1}$ in $\mathcal{R}[x]$ such that $C = \langle g_0, \gamma g_1, \ldots, \gamma^{e-1} g_{e-1}\rangle$ and $g_{e-1} \mid g_{e-2} \mid \cdots \mid g_1 \mid g_0 \mid (x^n-1)$. Then*

1. $\gamma^{e-1} g = \gamma^{e-1} g_{e-1} \frac{G}{\hat{g}_0}$,
2. $\gamma^{e-1-i}(\prod_{j=0}^{i-1} \hat{g}_j) g = \gamma^{e-1} g_{e-1} \frac{G}{\hat{g}_i}$, *for* $i = 1, \ldots, e-1$.

Proof. Let $g = g_0 + \gamma g_1 + \cdots + \gamma^{e-1} g_{e-1}$. Then

$$\gamma^{e-1} g = \gamma^{e-1} g_0 \tfrac{1}{g_1} \tfrac{g_1}{g_2} \cdots \tfrac{g_{e-3}}{g_{e-2}} \tfrac{g_{e-2}}{g_{e-1}} g_{e-1} = \gamma^{e-1} g_{e-1} \hat{g}_1 \hat{g}_2 \cdots \hat{g}_{e-2} \hat{g}_{e-1} = \gamma^{e-1} g_{e-1} \tfrac{G}{\hat{g}_0},$$

and 1 holds. For $i = 1, \ldots, e-1$ we have that

$$\begin{aligned}\gamma^{e-1-i}(\prod_{j=0}^{i-1}\hat{g}_j)g =& \gamma^{e-1-i}(\prod_{j=0}^{i-1}\hat{g}_j)\gamma^i g_i \frac{1}{g_{i+1}}\frac{g_{i+1}}{g_{i+2}}\cdots\frac{g_{e-2}}{g_{e-1}}g_{e-1}\\ =& \gamma^{e-1}g_{e-1}\hat{g}_0\hat{g}_1\ldots\hat{g}_{i-1}\hat{g}_{i+1}\ldots\hat{g}_{e-1} = \gamma^{e-1}g_{e-1}\frac{G}{\hat{g}_i},\end{aligned}$$

and statement 2 is proved.

□

Corollary 2.3 (of Th. 2.1)**.** *Let C be a cyclic code of length n over a finite chain ring $\mathcal{R}$ such that $C = \langle g_0, \gamma g_1, \ldots, \gamma^{e-1}g_{e-1}\rangle$ with $g_{e-1} \mid g_{e-2} \mid \cdots \mid g_1 \mid g_0 \mid (x^n-1)$. Then,*

$$|C| = |\mathcal{R}/\langle\gamma\rangle|^{\sum_{i=0}^{e-1}(e-i)\deg(\hat{g}_i)}.$$

Proof. From the previous definition of $\hat{g}_i$, these polynomials are the same polynomials described in [6, Theorem 3.4]. □

Theorem 2.4. *Let C be a cyclic code of length n over a finite chain ring $\mathcal{R}$, which has maximal ideal $\langle\gamma\rangle$ and e is the nilpotency of γ. Let $g_0, g_1, \ldots, g_{e-1}$ polynomials in $\mathcal{R}[x]$ such that $C = \langle g_0, \gamma g_1, \ldots, \gamma^{e-1}g_{e-1}\rangle$ and $g_{e-1} \mid g_{e-2} \mid \cdots \mid g_1 \mid g_0 \mid (x^n-1)$. Then the polynomial $g = g_0 + \gamma g_1 + \cdots + \gamma^{e-1}g_{e-1}$ is a generating polynomial of C, i.e., $C = \langle g\rangle$.*

Proof. Clearly $g = g_0+\gamma g_1+\cdots+\gamma^{e-1}g_{e-1} \in C$. then, we only have to prove that $\gamma^i g_i \in \langle g\rangle$, for all $i = 0, \ldots, e-1$.

Case $e = 2$: We have that $g = g_0 + \gamma g_1$, and $C = \langle g_0, \gamma g_1\rangle$. Then, $\gamma g = \gamma g_0 = \gamma g_1 \frac{g_0}{g_1}$ and $\frac{x^n-1}{g_0}g = \gamma g_1 \frac{x^n-1}{g_0}$, since $\frac{g_0}{g_1}$ and $\frac{x^n-1}{g_0}$ are coprime then γg_1 belongs to $\langle g\rangle$ and hence g_0 also belongs to $\langle g\rangle$.

Case $e = 3$: We have that $g = g_0 + \gamma g_1 + \gamma^2 g_2$, and $C = \langle g_0, \gamma g_1, \gamma^2 g_2\rangle$. Now $\gamma^2 g = \gamma^2 g_2 \frac{g_1}{g_2}\frac{g_0}{g_1}$, $\gamma\frac{x^n-1}{g_0}g = \gamma^2 g_2 \frac{x^n-1}{g_0}\frac{g_1}{g_2}$ and $\frac{x^n-1}{g_0}\frac{g_0}{g_1}g = \gamma^2 g_2 \frac{x^n-1}{g_0}\frac{g_0}{g_1}$, since $\gcd(\frac{g_0}{g_1}, \frac{g_1}{g_2}, \frac{x^n-1}{g_0}) = 1$ then $\gamma^2 g_2$ belongs to $\langle g\rangle$, hence $g_0+\gamma g_1$. So $\langle g\rangle = \langle g_0+\gamma g_1, \gamma^2 g_2\rangle$. Arguing as in case $e = 2$ it is straightforward that $\langle g\rangle = \langle g_0, \gamma g_1, \gamma^2 g_2\rangle$.

In the general case, let $g = g_0 + \gamma g_1 + \cdots + \gamma^{e-1}g_{e-1}$ and define, as in Lemma 2.2, the polynomials G and $\hat{g}_i$ for $i \in \{0, \ldots, e-1\}$. Then, $\gamma^{e-1}g = \gamma^{e-1}g_{e-1}\frac{G}{\hat{g}_0} \in \langle g\rangle$ and $\gamma^{e-1-i}(\prod_{j=0}^{i-1}\hat{g}_j)g = \gamma^{e-1}g_{e-1}\frac{G}{\hat{g}_i} \in \langle g\rangle$, for $i \in \{1, \ldots, e-1\}$. Since $\gcd(\frac{G}{\hat{g}_0}, \frac{G}{\hat{g}_1}, \ldots, \frac{G}{\hat{g}_{e-1}}) = 1$, we have that $\gamma^{e-1}g_{e-1} \in \langle g\rangle$ and $\langle g\rangle = \langle g_0 + \gamma g_1 + \cdots + \gamma^{e-2}g_{e-2}, \gamma^{e-1}g_{e-1}\rangle$.

Reasoned similarly, one obtains that $\langle g\rangle = \langle g_0, \gamma g_1, \ldots, \gamma^{e-1}g_{e-1}\rangle$. □

Theorem 2.5. *Let $C = \langle g\rangle = \langle g_0 + \gamma g_1 + \cdots + \gamma^{e-2}g_{e-2} + \gamma^{e-1}g_{e-1}\rangle$ be a cyclic code of length n over a finite chain ring $\mathcal{R}$, which has maximal ideal $\langle\gamma\rangle$ and e is the nilpotency of*

γ with $g_{e-1} \mid g_{e-2} \mid \cdots \mid g_1 \mid g_0 \mid (x^n-1)$. We define the sets

$$S_{\gamma^j} = \left[x^i (\prod_{t=0}^{j-1} \hat{g}_t) g \right]_{i=0}^{\deg(\hat{g}_j)},$$

for $0 \leq j < e$. Then,

$$S = \bigcup_{j=0}^{e-1} S_{\gamma^j}$$

forms a minimal generating set for C as an $\mathcal{R}$-module.

Proof. Let $c \in \mathcal{C}$. Then, $c = dg$ with $d \in \mathcal{R}[x]$. If $\deg(d) < \deg(\hat{g}_0)$ then $dg \in \langle S_{\gamma^0} \rangle_{\mathcal{R}}$ and $c \in \langle S \rangle_{\mathcal{R}}$. Otherwise, compute $d = d_0\hat{g}_0 + r_0$ with $\deg(r_0) < \deg(\hat{g}_0)$, so $dg = d_0\hat{g}_0 g + r_0 g$ and $r_0 g \in \langle S_{\gamma^0} \rangle_{\mathcal{R}}$.

If $\deg(d_0) < \deg(\hat{g}_1)$ then $d_0\hat{g}_0 g \in \langle S_{\gamma^1} \rangle_{\mathcal{R}}$ and $c \in \langle S \rangle_{\mathcal{R}}$. Otherwise, compute $d_0 = d_1\hat{g}_1 + r_1$ with $\deg(r_1) < \deg(\hat{g}_1)$, so $d_0\hat{g}_0 g = d_1\hat{g}_1\hat{g}_0 g + r_1\hat{g}_0 g$ and $r_1\hat{g}_0 g \in \langle S_{\gamma^1} \rangle_{\mathcal{R}}$.

In the worst case and reasoning similarly, one obtains that $c \in \langle S \rangle_{\mathcal{R}}$ if $d_{e-2}(\prod_{t=0}^{e-2} \hat{g}_t) g \in \langle S \rangle_{\mathcal{R}}$. It is obvious that if $\deg(d_{e-2}) < \deg(\hat{g}_{e-1})$ then $d_{e-2}(\prod_{t=0}^{e-2} \hat{g}_t) g \in \langle S_{\gamma^{e-1}} \rangle_{\mathcal{R}}$, if not, $d_{e-2} = d_{e-1}\hat{g}_{e-1} + r_{e-1}$. Therefore,

$$d_{e-2}(\prod_{t=0}^{e-2} \hat{g}_t) g = d_{e-1}(\prod_{t=0}^{e-1} \hat{g}_t) g + r_{e-1}(\prod_{t=0}^{e-2} \hat{g}_t) g = r_{e-1}(\prod_{t=0}^{e-2} \hat{g}_t) g \in \langle S_{\gamma^{e-1}} \rangle.$$

Since $r_{e-1}(\prod_{t=0}^{e-2} \hat{g}_t) g \in \langle S_{\gamma^{e-1}} \rangle$ then $c \in \langle S \rangle_{\mathcal{R}}$, so S is a generating set. By the definition of S clearly

$$|S| = |\mathcal{R}/\langle\gamma\rangle|^{\sum_{i=0}^{e-1}(e-i)\deg(\hat{g}_i)}.$$

By Corollary 2.3, $|C| = |S|$ and S is a minimal generating set. □

3 Linear codes over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$

Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be finite chain rings where γ_1 and γ_2 are generators of the maximal ideals of $\mathcal{R}_1$ and $\mathcal{R}_2$ with nilpotency indices e_1 and e_2, respectively. We will suppose that $\mathcal{R}_1$ and $\mathcal{R}_2$ have the same residue field $K = \mathcal{R}_1/\langle\gamma_1\rangle = \mathcal{R}_2/\langle\gamma_2\rangle$, with $|K| = q = p^m$. By $^-\colon \mathcal{R}_i \to K$, we will denote the natural projection that maps $r \mapsto \bar{r} = r + \langle\gamma_i\rangle$, for $i = 1$ or 2.

Let $T_1 = \{r_0, \ldots, r_{q-1}\}$ and $T_2 = \{r'_0, \ldots, r'_{q-1}\}$ be the Teichmüller sets of representatives of $\mathcal{R}_1$ and $\mathcal{R}_2$, resp., then we can arrange the subscripts such that $\bar{r}_i = \bar{r'}_i$. Assume that $e_1 \leq e_2$. Then we can consider the surjective ring homomorphism

$$\begin{array}{rccc} \pi: & \mathcal{R}_2 & \to & \mathcal{R}_1 \\ & \gamma_2 & \mapsto & \gamma_1 \\ & r'_j & \mapsto & r_j. \end{array}$$

 ISBN: 978-84-608-6082-2

Note that $\pi(\gamma_2{}^i) = 0$ if $i \geq e_1$. For $a \in \mathcal{R}_2$ and $b \in \mathcal{R}_1$, define a multiplication $*$ as follows: $a * b = \pi(a)b$. Then, $\mathcal{R}_1$ is an $\mathcal{R}_2$-module with external multiplication $*$ given by π. Since $\mathcal{R}_1$ is commutative then $*$ has the commutative property. Then, we can generalize this multiplication over the ring $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ as follows. Let a be an element of $\mathcal{R}_2$ and $\mathbf{u} = (u \mid u') = (u_0, u_1, \ldots, u_{\alpha-1} \mid u'_0, u'_1, \ldots, u'_{\beta-1}) \in \mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$. Then,

$$a*\mathbf{u} = (\pi(a)u_0, \pi(a)u_1, \ldots, \pi(a)u_{\alpha-1} \mid au'_0, au'_1, \ldots, au'_{\beta-1}).$$

With this external operation the ring $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ is also an $\mathcal{R}_2$-module.

Definition 3.1. *A subset $\mathcal{C} \subseteq \mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ is a linear code if it is a submodule of $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$.*

The next result gives the structure of a generator matrix of a linear code.

Proposition 3.2. *Let $\mathcal{C}$ be a linear code over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$. Then $\mathcal{C}$ is permutation equivalent to a code with generator matrix of the form*

$$G = \left(\begin{array}{c|c} B & T \\ S & A \end{array}\right),$$

where

$$B = \begin{pmatrix} I_{k_0} & B_{0,1} & B_{0,2} & B_{0,3} & \cdots & B_{0,e_1-1} & B_{0,e_1} \\ 0 & \gamma_1 I_{k_1} & \gamma_1 B_{1,2} & \gamma_1 B_{1,3} & \cdots & \gamma_1 B_{1,e_1-1} & \gamma_1 B_{1,e_1} \\ 0 & 0 & \gamma_1^2 I_{k_2} & \gamma_1^2 B_{2,3} & \cdots & \gamma_1^2 B_{2,e_1-1} & \gamma_1^2 B_{2,e_1} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \gamma_1^{e_1-1} I_{k_{e_1-1}} & \gamma_1^{e_1-1} B_{e_1-1,e_1} \end{pmatrix},$$

$$T = \begin{pmatrix} 0 \cdots & \gamma_2^{e_2-e_1} T_{0,1} & \gamma_2^{e_2-e_1} T_{0,2} & \cdots & \gamma_2^{e_2-e_1} T_{0,e_1} \\ 0 \cdots & 0 & \gamma_2^{e_2-e_1+1} T_{1,2} & \cdots & \gamma_2^{e_2-e_1+1} T_{1,e_1} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 \cdots & 0 & 0 & \cdots & \gamma_2^{e_2-1} T_{e_1-1,e_1} \end{pmatrix},$$

$$A = \begin{pmatrix} I_{l_0} & A_{0,1} & A_{0,2} & A_{0,3} & \cdots & A_{0,e_1-1} & A_{0,e_2} \\ 0 & \gamma_2 I_{l_1} & \gamma_2 A_{1,2} & \gamma_2 A_{1,3} & \cdots & \gamma_2 A_{1,e_2-1} & \gamma_2 A_{1,e_2} \\ 0 & 0 & \gamma_2^2 I_{l_2} & \gamma_2^2 A_{2,3} & \cdots & \gamma_2^2 A_{2,e_2-1} & \gamma_2^2 A_{2,e_2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \gamma_2^{e_2-1} I_{l_{e_2-1}} & \gamma_2^{e_2-1} A_{e_2-1,e_2} \end{pmatrix},$$

$$S = \begin{pmatrix} 0 & S_{0,1} & S_{0,2} & \cdots & S_{0,e_1-1} & S_{0,e_1} \\ 0 & S_{1,1} & S_{1,2} & \cdots & S_{1,e_1-1} & S_{1,e_1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & S_{e_2-e_1-1,1} & S_{e_2-e_1-1,2} & \cdots & S_{e_2-e_1-1,e_1-1} & S_{e_2-e_1-1,e_1} \\ 0 & 0 & \gamma_1 S_{e_2-e_1,2} & \cdots & \gamma_1 S_{e_2-e_1,e_1-1} & \gamma_1 S_{e_2-e_1,e_1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \gamma_1^{e_1-2} S_{e_2-3,e_1-1} & \gamma_1^{e_1-3} S_{e_2-2,e_1} \\ 0 & 0 & 0 & \cdots & 0 & \gamma_1^{e_1-1} S_{e_2-2,e_1} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

where the entries in $\gamma_1^i B_{i,j}$ and $\gamma_1^i S_{i,j}$ are in $\langle \gamma_1^i \rangle$ and the ones in $\gamma_2^t A_{t,j}$ and $\gamma_2^t T_{t,j}$ are in $\langle \gamma_2^t \rangle$.

Proof. Similiar to [2, Theorem 4] □

Let $\mathcal{C}_X$ be the canonical projection of $\mathcal{C}$ on the first α coordinates and $\mathcal{C}_Y$ on the last β coordinates. The canonical projection is a linear map. Then, $\mathcal{C}_X$ and $\mathcal{C}_Y$ are $\mathcal{R}_1$ and $\mathcal{R}_2$ linear codes of length α and β, respectively. A code $\mathcal{C}$ is called *separable* if $\mathcal{C}$ is the direct product of $\mathcal{C}_X$ and $\mathcal{C}_Y$, i.e., $\mathcal{C} = \mathcal{C}_X \times \mathcal{C}_Y$. Moreover, if $\mathcal{C}$ is separable then

$$G = \left(\begin{array}{c|c} B & 0 \\ 0 & A \end{array}\right).$$

Example 1. *Let $\mathcal{C}$ be a linear code over $\mathbb{Z}_2^3 \times \left(\frac{\mathbb{Z}_2[u]}{\langle u^3\rangle}\right)^4$ generated by the matrix*

$$\left(\begin{array}{ccc|cccc} 1 & 1 & 0 & u & u+u^2 & 1+u & 1+u^2 \\ 0 & 1 & 0 & 1 & u & u^2 & 0 \\ 0 & 1 & 1 & 0 & u^2 & 0 & u^2 \\ 1 & 1 & 1 & u^2 & u & u+u^2 & 0 \end{array}\right).$$

Hence, as described in Theorem 3.2, $\mathcal{C}$ is permutation equivalent to a code generated by the following matrix:

$$G = \left(\begin{array}{ccc|ccc|c} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 & 0 & u \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & u & u+u^2 \end{array}\right).$$

Note that $\mathcal{C}$ is not separable.

4 Cyclic codes over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$

Definition 4.1. *Let $\mathcal{C}$ be a linear code over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$. The code $\mathcal{C}$ is called* cyclic *if*

$$(u_0, u_1, \ldots, u_{\alpha-2}, u_{\alpha-1} \mid u'_0, u'_1, \ldots, u'_{\beta-2}, u'_{\beta-1}) \in \mathcal{C}$$

implies

$$(u_{\alpha-1}, u_0, u_1, \ldots, u_{\alpha-2} \mid u'_{\beta-1}, u'_0, u'_1, \ldots, u'_{\beta-2}) \in \mathcal{C}.$$

Let $\mathbf{u} = (u_0, u_1, \ldots, u_{\alpha-1} \mid u'_0, \ldots, u'_{\beta-1})$ be a codeword in $\mathcal{C}$ and let i be an integer. We then denote by $\mathbf{u}^{(i)} = (u_{0+i}, u_{1+i}, \ldots, u_{\alpha-1+i} \mid u'_{0+i}, \ldots, u'_{\beta-1+i})$ the ith shift of $\mathbf{u}$, where the subscripts are read modulo α and β, respectively.

We remark that in this paper the definition of a cyclic code over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ is clear as long as $\mathcal{R}_1$ and $\mathcal{R}_2$ are different rings, since the elements on the first α coordinates and the ones in the last β coordinates belong from different rings, $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. In the

particular case that $\mathcal{R}_1 = \mathcal{R}_2$, the cyclic code over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ is known in the literature as *double cyclic code*, see [4], [7]. The term double cyclic is given in order to distinguish the cyclic code over $\mathcal{R}_1^\alpha \times \mathcal{R}_1^\beta$ from the cyclic code over $\mathcal{R}_1^{\alpha+\beta}$.

Note that $\mathcal{C}_X$ and $\mathcal{C}_Y$ are $\mathcal{R}_1$ and $\mathcal{R}_2$ cyclic codes of length α and β, respectively.

Denote by $\mathcal{R}_{\alpha,\beta}$ the ring $\mathcal{R}_1[x]/(x^\alpha - 1) \times \mathcal{R}_2[x]/(x^\beta - 1)$. There is a bijective map between $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ and $\mathcal{R}_{\alpha,\beta}$ where $\mathbf{u} = (u_0, u_1, \ldots, u_{\alpha-1} \mid u'_0, \ldots, u'_{\beta-1})$ maps to $\mathbf{u}(x) = (u_0 + u_1 x + \cdots + u_{\alpha-1}x^{\alpha-1} \mid u'_0 + \cdots + u'_{\beta-1}x^{\beta-1})$.

Note that we can extend the map π to the polynomial ring $\mathcal{R}_2[x]$ applying this map to each of the coefficients of a given polynomial.

Definition 4.2. *Define the operation* $* : \mathcal{R}_2[x] \times \mathcal{R}_{\alpha,\beta} \to \mathcal{R}_{\alpha,\beta}$ *as*

$$\lambda(x) * (p(x) \mid q(x)) = (\pi(\lambda(x))p(x) \mid \lambda(x)q(x)),$$

where $\lambda(x) \in \mathcal{R}_2[x]$ *and* $(p(x) \mid q(x)) \in \mathcal{R}_{\alpha,\beta}$.

The ring $\mathcal{R}_{\alpha,\beta}$ with the external operation $*$ is a $\mathcal{R}_2[x]$-module. Let $\mathbf{u}(x) = (u(x) \mid u'(x))$ be an element of $\mathcal{R}_{\alpha,\beta}$. Note that if we operate $\mathbf{u}(x)$ by x we get

$$\begin{aligned} x * \mathbf{u}(x) = x * (u(x) \mid u'(x)) &= (u_0 x + \cdots + u_{\alpha-1}x^\alpha \mid u'_0 x + \cdots + u'_{\beta-1}x^\beta) \\ &= (u_{\alpha-1} + \cdots + u_{\alpha-2}x^{\alpha-1} \mid u'_{\beta-1} + \cdots + u'_{\beta-2}x^{\beta-1}). \end{aligned}$$

Hence, $x * \mathbf{u}(x)$ is the image of the vector $\mathbf{u}^{(1)}$. Thus, the operation of $\mathbf{u}(x)$ by x in $R_{\alpha,\beta}$ corresponds to a shift of $\mathbf{u}$. In general, $x^i * \mathbf{u}(x) = \mathbf{u}^{(i)}(x)$ for all i.

4.1 Algebraic structure and generators of cyclic codes

In this subsection, we study submodules of $\mathcal{R}_{\alpha,\beta}$. We describe the generators of such submodules and state some properties. From now on, $\langle S \rangle$ will denote the $\mathcal{R}_2[x]$-submodule generated by a subset S of $\mathcal{R}_{\alpha,\beta}$.

For the rest of the discussion we will consider that α and β are coprime integers with p. From this assumption, we know that $\mathcal{R}_1[x]/\langle x^\alpha - 1\rangle$ and $\mathcal{R}_2[x]/\langle x^\beta - 1\rangle$ are principal ideal rings, see [6].

Theorem 4.3. *Every submodule* $\mathcal{C}$ *of the* $\mathcal{R}_2[x]$*-module* $\mathcal{R}_{\alpha,\beta}$ *can be written as*

$$\mathcal{C} = \langle (b(x) \mid 0), (\ell(x) \mid a(x)) \rangle,$$

where $b(x), a(x)$ *are generator polynomials in* $\mathcal{R}_1[x]/(x^\alpha - 1)$ *and* $\mathcal{R}_2[x]/(x^\beta - 1)$ *resp., and* $\ell(x) \in \mathcal{R}_1[x]/(x^\alpha - 1)$.

Proof. Let $\psi_X : \mathcal{R}_{\alpha,\beta} \to \mathcal{R}_1[x]/(x^\alpha - 1)$ and $\psi_Y : \mathcal{R}_{\alpha,\beta} \to \mathcal{R}_2[x]/(x^\beta - 1)$ be the canonical projections, let $\mathcal{C}$ be a submodule of $\mathcal{R}_{\alpha,\beta}$.
Define $\mathcal{C}' = \{(p(x)|q(x)) \in \mathcal{C} \mid q(x) = 0\}$. It is easy to check that $\mathcal{C}' \cong \psi_X(\mathcal{C}')$ by $(p(x) \mid 0) \mapsto p(x)$. Hence, by Theorem 2.4, $\psi_X(\mathcal{C}')$ is finitely generated by one element and so is $\mathcal{C}'$. Let $b(x)$ be a generator of $\psi_X(\mathcal{C}')$, then $(b(x) \mid 0)$ is a generator of $\mathcal{C}'$.
As $\mathcal{R}_2[x]/(x^\beta - 1)$ is also a principal ideal ring, then $\mathcal{C}_Y = \psi_Y(\mathcal{C})$ is generated by one element. Let $a(x) \in \mathcal{C}_Y$ such that $\mathcal{C}_Y = \langle a(x) \rangle$, then there exists $\ell(x) \in \mathcal{R}_1[x]/(x^\alpha - 1)$ such that $(\ell(x) \mid a(x)) \in \mathcal{C}$.
We claim that

$$\mathcal{C} = \langle (b(x) \mid 0), (\ell(x) \mid a(x)) \rangle.$$

Let $(p(x) \mid q(x)) \in \mathcal{C}$, then $q(x) = \psi_Y(p(x) \mid q(x)) \in \mathcal{C}_Y$. So, there exists $\lambda(x) \in \mathcal{R}_2[x]$ such that $q(x) = \lambda(x)a(x)$. Now,

$$(p(x) \mid q(x)) - \lambda(x) * (\ell(x) \mid a(x)) = (p(x) - \pi(\lambda(x))\ell(x) \mid 0)$$

belongs to $\mathcal{C}'$. Then, there exists $\mu(x) \in \mathcal{R}_2[x]$ such that $(p(x) - \pi(\lambda(x))\ell(x) \mid 0) = \mu(x) * (b(x) \mid 0)$. Thus,

$$(p(x) \mid q(x)) = \mu(x) * (b(x) \mid 0) + \lambda(x) * (\ell(x) \mid a(x)).$$

So, $\mathcal{C}$ is finitely generated by $\langle (b(x) \mid 0), (\ell(x) \mid a(x)) \rangle$. □

From the previous results, it is clear that we can identify double cyclic codes in $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ as submodules of $\mathcal{R}_{\alpha,\beta}$. So, any submodule of $\mathcal{R}_{\alpha,\beta}$ is a cyclic code. From now on, we will denote by $\mathcal{C}$ indistinctly both the code and the corresponding submodule of $\mathcal{R}_{\alpha,\beta}$.

In the following, a polynomial $f(x)$ in $\mathcal{R}_1[x]$ or $\mathcal{R}_2[x]$ will be denoted simply by f.

Proposition 4.4. *Let $\mathcal{C}$ be a cyclic code over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$. Then, there exist polynomials ℓ and $b_{e_1-1}|b_{e_1-2}|\dots|b_1|b_0|(x^\alpha - 1)$ over $\mathcal{R}_1[x]$ and $a_{e_2-1}|a_{e_2-2}|\dots|a_1|a_0|(x^\beta - 1)$ over $\mathcal{R}_2[x]$ such that*

$$\mathcal{C} = \langle (b_0 + \gamma_1 b_1 + \cdots + {\gamma_1}^{e_1-1} b_{e_1-1} \mid 0), (\ell \mid a_0 + \gamma_2 a_1 + \cdots + {\gamma_2}^{e_2-1} a_{e_2-1}) \rangle.$$

Proof. Let $\mathcal{C}$ be a cyclic code over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$. By Theorem 4.3, there exist polynomials $b, \ell \in \mathcal{R}_1[x]/(x^\alpha - 1)$ and $a \in \mathcal{R}_2[x]/(x^\beta - 1)$ such that $\mathcal{C} = \langle (b \mid 0), (\ell \mid a) \rangle$. By Theorem 2.4, one can consider $b = b_0 + \gamma_1 b_1 + \cdots + {\gamma_1}^{e_1-1} b_{e_1-1}$ and $a = a_0 + \gamma_2 a_1 + \cdots + {\gamma_2}^{e_2-1} a_{e_2-1}$ such that $b_{e_1-1}|b_{e_1-2}|\dots|b_1|b_0|(x^\alpha - 1)$ and $a_{e_2-1}|a_{e_2-2}|\dots|a_1|a_0|(x^\beta - 1)$. □

For the rest of the discussion, any cyclic code $\mathcal{C}$ over $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ is of the form $\mathcal{C} = \langle (b \mid 0), (\ell \mid a) \rangle$, where $b = b_0 + \gamma_1 b_1 + \cdots + {\gamma_1}^{e_1-1} b_{e_1-1}$ and $a(x) = a_0 + \gamma_2 a_1 + \cdots + {\gamma_2}^{e_2-1} a_{e_2-1}$, for polynomials b_i and a_j as in Proposition 4.4.

Example 2. *Let* $\mathcal{R}_1 = \frac{\mathbb{F}_9[u]}{\langle u^2\rangle}$ *and* $\mathcal{R}_2 = \frac{\mathbb{F}_9[u]}{\langle u^3\rangle}$, *with* $\gamma_1 = u$, $e_1 = 2$, $\gamma_2 = u$, *and* $e_2 = 3$. *Let* ξ *be a generator of the multiplicative group* $\mathbb{F}_9^*$. *Consider the cyclic code over* $\left(\frac{\mathbb{F}_9[u]}{\langle u^2\rangle}\right)^4 \times \left(\frac{\mathbb{F}_9[u]}{\langle u^3\rangle}\right)^{10}$ *generated by*

$$\mathcal{C} = \langle (u(x^2-1) \mid 0), (u \mid (x^4 + \xi^3 x^2 + 1) + u(x^2 + \xi^5 x + 1) + u^2)\rangle.$$

Then,

$$b_0 = x^4 - 1, \quad b_1 = x^2 - 1, \quad \ell = u, a_0 = x^4 + \xi^3 x^2 + 1, \quad a_1 = x^2 + \xi^5 x + 1, \quad a_2 = 1.$$

4.2 Minimal generating sets

Our goal is to find a set generators for a cyclic code, $\mathcal{C}$, as an $\mathcal{R}_2$-module. Once we found it, we are going to use it to determine the size of $\mathcal{C}$ in terms of the generator polynomials.

Since b_0 is a factor of $x^\alpha - 1$ and for $i = 1 \dots e_1 - 1$ the polynomial b_i is a factor of b_{i-1}, we will denote $\hat{b}_0 = \frac{x^\alpha - 1}{b_0}$, $\hat{b}_i = \frac{b_{i-1}}{b_i}$ for $i = 1 \dots e_1 - 1$, and $\hat{b}_{e_1} = b_{e_1-1}$. In the same way, we define $\hat{a}_0 = \frac{x^\beta - 1}{a_0}$, $\hat{a}_j = \frac{a_{j-1}}{a_j}$ for $j = 1 \dots e_2 - 1$, and $\hat{a}_{e_2} = a_{e_2-1}$.

Theorem 4.5. *Let* $\mathcal{C}$ *be a cyclic code over* $\mathcal{R}_1^\alpha \times \mathcal{R}_2^\beta$ *which has maximals ideals* $\langle\gamma_1\rangle \subset \mathcal{R}_1$ *and* $\langle\gamma_2\rangle \subset \mathcal{R}_2$ *with nilpotent indices* e_1 *and* e_2, *respectively. Define*

$$B_j = \left[x^i \left(\prod_{t=0}^{j-1} \hat{b}_t\right) * (b \mid 0) \right]_{i=0}^{\deg(\hat{b}_j)-1},$$

for $0 \le j < e_1$, *and*

$$A_k = \left[x^i \left(\prod_{t=0}^{k-1} \hat{a}_t\right) * (\ell \mid a) \right]_{i=0}^{\deg(\hat{a}_k)-1}.$$

for $0 \le k < e_2$. *Then,*

$$S = \left(\bigcup_{j=0}^{e_1-1} B_j \right) \bigcup \left(\bigcup_{k=0}^{e_2-1} A_k \right)$$

forms a minimal generating set for $\mathcal{C}$ *as an* $\mathcal{R}_2$*-module. Moreover,*

$$|\mathcal{C}| = q^{\sum_{i=0}^{e_1-1}(e_1-i)\deg(\hat{b}_i) + \sum_{j=0}^{e_2-1}(e_2-j)\deg\hat{a}_j},$$

where q *is the cardinality of the residue field.*

Proof. By Theorem 2.5, it is clear that the elements in S are $\mathcal{R}_2$-lineal independent since $\left(\bigcup_{j=0}^{e_1-1} B_j\right)_X$ and $\left(\bigcup_{k=0}^{e_2-1} A_k\right)_Y$ are minimal generating sets for the codes $\mathcal{C}_X$ and $\mathcal{C}_Y$, respectively. Let c be a codeword of $\mathcal{C}$, then $c = q*(b \mid 0) + d*(\ell \mid a)$. Reasoning similarly as in Theorem 2.5, we have that $q*(b \mid 0) \in \langle \bigcup_{j=0}^{e_1-1} B_j \rangle_{\mathcal{R}_2}$. So, we have to prove that $d*(\ell \mid a) \in \langle S \rangle_{\mathcal{R}_2}$.

If $\deg(d) < \deg(\hat{a}_0)$ then $d*(\ell \mid a) \in \langle A_0 \rangle_{\mathcal{R}_2}$ and $c \in \langle S \rangle_{\mathcal{R}_2}$. Otherwise, compute $d = d_0\hat{a}_0 + r_0$ with $\deg(r_0) < \deg(\hat{a}_0)$, so $d*(\ell \mid a) = d_0\hat{a}_0 * (\ell \mid a) + r_0 * (\ell \mid a)$ and $r_0 * (\ell \mid a) \in \langle A_0 \rangle_{\mathcal{R}_2}$.

In the worst case and reasoning similarly, one obtains that $c \in \langle S \rangle_{\mathcal{R}_2}$ if $d_{e_2-2}(\prod_{t=0}^{e_2-2} \hat{a}_t) * (\ell \mid a) \in \langle S \rangle_{\mathcal{R}_2}$. It is obvious that if $\deg(d_{e_2-2}) < \deg(\hat{a}_{e_2-1})$ then $d_{e_2-2}(\prod_{t=0}^{e_2-2} \hat{a}_t) * (\ell \mid a) \in \langle A_{e_2-1} \rangle_{\mathcal{R}_2}$, if not $d_{e_2-2} = d_{e_2-1}\hat{a}_{e_2-1} + r_{e_2-1}$. Therefore,

$$d_{e_2-2}\left(\prod_{t=0}^{e_2-2} \hat{a}_t\right) * (\ell \mid a) = d_{e_2-1}\left(\prod_{t=0}^{e_2-1} \hat{a}_t\right) * (\ell \mid a) + r_{e_2-1}\left(\prod_{t=0}^{e_2-2} \hat{a}_t\right) * (\ell \mid a).$$

On one hand, $r_{e_2-1}(\prod_{t=0}^{e_2-2} \hat{a}_t) * (\ell \mid a) \in \langle A_{e_2-1} \rangle_{\mathcal{R}_2}$. On the other hand, $d_{e_2-1}(\prod_{t=0}^{e_2-1} \hat{a}_t) * (\ell \mid a) = d_{e_2-1}(\prod_{t=0}^{e_2-1} \hat{a}_t) * (\ell \mid 0)$ and then $d_{e_2-1}(\prod_{t=0}^{e_2-1} \hat{a}_t) * (\ell \mid a) = f * (b \mid 0) \in \langle \bigcup_{j=0}^{e_1-1} B_j \rangle_{\mathcal{R}_2}$. Thus, $c \in \langle S \rangle_{\mathcal{R}_2}$ and S is a minimal generating set for $\mathcal{C}$. □

Example 3. *Consider* $\mathcal{R}_{\alpha,\beta} = \mathbb{Z}_2[x]/(x^2-1) \times \mathbb{Z}_4[x]/(x^3-1)$ *and the cyclic code*

$$\mathcal{C} = \langle (x-1 \mid (x^2+x+1)+2) \rangle,$$

where $b_0(x) = x^2-1$, $\ell(x) = x-1$, $a_0(x) = x^2+x+1$ *and* $a_1(x) = 1$. *Then,* $S = \{(x-1 \mid x^2+x+3), (0 \mid 2x+2), (0 \mid 2x^2+2x)\}$ *and*

$$|\mathcal{C}| = 2^{\sum_{j=0}^{2-1}(2-j)deg(\hat{a}_j)} = 2^4 = 16.$$

Example 4. *From Example 2, consider the cyclic code over* $\left(\frac{\mathbb{F}_9[u]}{\langle u^2 \rangle}\right)^4 \times \left(\frac{\mathbb{F}_9[u]}{\langle u^3 \rangle}\right)^{10}$ *generated by*

$$\mathcal{C} = \langle (u(x^2-1) \mid 0), (u \mid (x^4+\xi^3x^2+1) + u(x^2+\xi^7x+1) + u^2) \rangle.$$

Let $b = u(x^2-1), \ell = u$ *and* $a = (x^4+\xi^3x^2+1) + u(x^2+\xi^7x+1) + u^2$. *Then, a minimal generating set for* $\mathcal{C}$ *is the union of*

$$B_0 = \emptyset,\ B_1 = \left[x^i * (u(x^2-1) \mid 0)\right]_{i=0}^{1},$$

$$A_0 = \left[x^i * (\ell \mid a)\right]_{i=0}^{5},\ A_1 = \left[x^i\mu * (\ell \mid a)\right]_{i=0}^{1},$$

and

$$A_2 = \left[x^i\lambda * (\ell \mid a)\right]_{i=0}^{1},$$

where $\mu = x^6 + \xi^7x^4 + \xi^3x^2 + 2$ *and* $\lambda = x^8 + \xi x^7 + \xi x^6 + x^5 + 2x^3 + \xi^5x^2 + \xi^5x + 2$. *So,*

$$|\mathcal{C}| = 9^{\sum_{i=0}^{2-1}(2-i)deg(\hat{b}_i) + \sum_{j=0}^{3-1}(3-j)deg(\hat{a}_j)} = 9^{2+24} = 9^{26}.$$

 ISBN: 978-84-608-6082-2

Acknowledgements

This work has been partially supported by the Spanish MINECO grants TIN2013-40524-P and MTM2015-69138-REDT, and by the Catalan AGAUR grant 2014SGR-691.

References

[1] T. Abualrub, I. Siap, N. Aydin, $\mathbb{Z}_2\mathbb{Z}_4$-*additive cyclic codes*, IEEE Trans. Info. Theory **60(3)** (2014) 1508–1514.

[2] I. Aydogdu, I. Siap, *On* $\mathbb{Z}_{p^r}\mathbb{Z}_{p^s}$*-additive codes*, Linear and Multilinear Algebra **63(10)** (2014) 2089–2102.

[3] J. Borges, C. Fernández-Córdoba, J. Pujol, J. Rifà and M. Villanueva, $\mathbb{Z}_2\mathbb{Z}_4$*-linear codes: generator matrices and duality*, Des., Codes and Crypto. **54(2)** (2010) 167–179.

[4] J. Borges, C. Fernández-Córdoba, R. Ten-Valls, $\mathbb{Z}_2$*-double cyclic codes*, arXiv:1410.5604.

[5] J. Borges, C. Fernández-Córdoba, R. Ten-Valls, $\mathbb{Z}_2\mathbb{Z}_4$*-additive cyclic codes, generator polynomials and dual codes*, arXiv:1406.4425.

[6] H. Q. Dinh, S. R. López-Permouth, *Cyclic and negacyclic codes over finite chain rings*, IEEE Trans. Info. Theory **50(8)** (2004) 1728–1744.

[7] J. Gao, M. Shi, T. Wu and F. Fu, *On double cyclic codes over* $\mathbb{Z}_4$, Finite Fields and Their Applications **39** (2016) 233–250.

[8] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, P. Solé, *The* $\mathbb{Z}_4$*-Linearity of Kerdock, Preparata, Goethals, and Related Codes*, IEEE Trans. Info. Theory **40(2)** (1994) 301–319.

[9] G. Norton, A. Salagean, *On the structure of linear and cyclic codes over finite chain rings* Appl. Algebra Eng. Commun. Comput. **10** (2000) 489–506.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Competition between algae and fungi in a lake: a mathematical model

Iulia Martina Bulai[1] **and Ezio Venturino**[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino, via Carlo Alberto 10, 10123 Torino, Italy*

emails: iuliam@live.it, ezio.venturino@unito.it

Abstract

In this paper a mathematical model for handling water pollution is introduced. We assume that algae and fungi are in competition for resources that come from wastewater. Both algae and fungi need dissolved oxygen (DO) for their biologial process of growth. But there is a difference, indeed algae produce it too and in a higher quantity than the one they use. It is shown that if the coexistence equilibrium exists, it is stable without additional conditions. If the competition rate between algae and fungi is not high for a chosen set of parameters the stability of the coexistence equilibrium is reached even without an external constant input of DO in the system.

Key words: mathematical model, algae, fungi, competition, wastewater
MSC 2000: AMS codes (optional)

1 Introduction

Algae are important in a lake, they can improve the quality of the aquatic ecosystem. Under right conditions such as adequate nutrients (mostly phosphorus, but nitrogen is important too) they grow. The nutrients that are present in the wastewater can derive from agricultural and/or industrial discharges. Fungi can be used for biodegradation of organic pollutant in a waterbody, and they grow using the nutrients obtained from the biodegradation, [1]. Some mathematical models in the literature study the behavior of algae biomass in a waterbody in the presence of organic pollutants, [4, 5]. In [2, 3] the case of fungi has been addressed. In this paper we want to study what happens when both algae and fungi are present in the same waterbody, for example in a lake. Furthermore we suppose that they are in competition for the resources coming from the pollutants.

 ISBN: 978-84-608-6082-2

2 The mathematical model

In this paper we introduce a mathematical system that models the behavior of algae and fungi in a waterbody. The waterbody considered could be nutrient-rich waters, like municipal wastewater or some industrial effluents. Both algae and fungi can feed on these wastes and therefore purify the water, while also producing a biomass suitable for biofuels production. Thus algae and fungi are in competition for food, since both share the same resources. Further, fungi as well as algae need DO to thrive but we assume that the algae's production and input of DO into the system is much larger than their own use for their growth.

The model consists of three equations that describe the time evolution of the algae population, the fungi population and the DO respectively. The model, in which all the parameters are nonnegative, reads:

$$\begin{aligned} \frac{dA}{dt} &= r_A A - a_A A - b_A A^2 - cAF \\ \frac{dF}{dt} &= \frac{hOF}{k + k_O O} - a_F F - b_F F^2 - cAF \\ \frac{dO}{dt} &= q_O + gA - a_O O - f\frac{hOF}{k + k_O O}. \end{aligned} \tag{1}$$

In the first equation algae grow at a constant rate r_A and are washed out at a constant rate a_A. We assume that algae are in competition among themselves at a constant rate b_A and also experience interspecific competition with fungi at rate c.

In the second equation the fungi's growth depends on the presence of DO. They are washed out at rate a_F. The intraspecific competition occurs at rate b_F while c denotes the rate of the interspecific competition with the algae population.

The third equation shows the evolution in time of DO. We assume that it is supplied from external sources at rate q_O, but a part of it comes from the algae own production at rate g. We take further into account its washing out, at rate a_O and its depletion due to its assumption by fungi at rate $f \geqslant 1$.

3 The qualitative analysis of the model

First of all to find the equilibrium points of the model, we need to solve the system obtained by setting the right hand side of (1) to zero,

$$\begin{cases} A(r_A - a_A - b_A A - cF) = 0 \\ F\left(\dfrac{hO}{k + k_O O} - a_F - b_F F - cA\right) = 0 \\ q_O + gA - a_O O - f\dfrac{hOF}{k + k_O O} = 0. \end{cases} \tag{2}$$

Further, for the stability analysis, we need to calculate the Jacobian matrix of the system (1)

$$J = \begin{bmatrix} r_A - a_A - 2b_A A - cF & -cA & 0 \\ -cF & -a_F - 2b_F F - cA + \dfrac{hO}{k + k_O O} & \dfrac{hkF}{(k + k_O O)^2} \\ g & -\dfrac{fhO}{k + k_O O} & -a_O - \dfrac{fhkF}{(k + k_O O)^2} \end{bmatrix}. \tag{3}$$

Solving (2) we obtain the analytic expression of three equilibrium points. In addition, we prove that two other equilibria exist. We also show that all these points are conditionally locally asymptotically stable, while the coexistence equilibrium is stable if it is feasible.

Proposition 1. The trivial equilibrium point, $E_0 = (0, 0, 0)$, exists if

$$q_O = 0. \tag{4}$$

Furthermore, it is stable if the following condition holds:

$$r_A < a_A. \tag{5}$$

Proof. For $A = F = O = 0$ in the system (2) we get that E_0 exists if $q_O = 0$. The characteristic polynomial associated to the matrix (3) evaluated at E_0 is

$$\det(J - \mu) = (r_A - a_A - \mu)(-a_F - \mu)(-a_O - \mu) = 0.$$

To have the stability of E_0 all the eigenvalues shoud be negative thus the condition (5) must hold.

□

Proposition 2. The fungi-and-algae-free point $E_1 = \left(0, 0, q_O a_O^{-1}\right)$ exist always. It is stable if the following conditions hold:

$$r_A < a_A \quad \text{and} \quad \frac{hq_O}{ka_O + k_O q_O} < a_F. \tag{6}$$

Proof. In fact for $A = F = 0$ in (2) from the last equation we get $O = q_O a_O^{-1}$. While the characteristic polynomial associated to E_1 is

$$\det(J - \mu) = (r_A - a_A - \mu)(-a_O - \mu)\left(\frac{hq_O}{ka_O + k_O q_O} - a_F - \mu\right) = 0.$$

To have all the eigenvalues negative the conditions (6) must hold. □

 ISBN: 978-84-608-6082-2

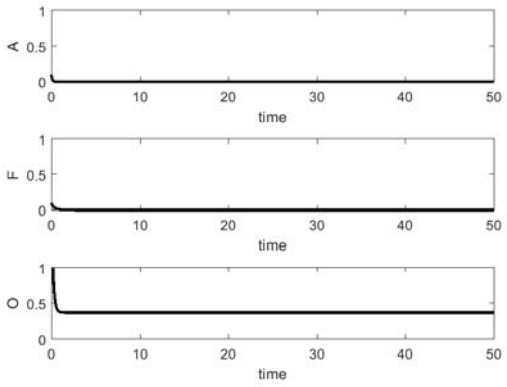


Figure 1: The equilibrium E_1 is stably achieved with the parameter values $r_A = 10.1273$, $a_A = 16.93072$, $b_A = 11.7382$, $c = 19.3012$, $h = 1.61592$, $k = 0.454245$, $k_O = 5.87845$, $a_F = 2.55798$, $b_F = 0.0344478$, $q_O = 2$, $g = 3.63317$, $a_O = 5.41771$, $f = 1$.

In Figure 1 one can see that for a chosen set of parameters the equilibrium E_1 is stably achieved.

Proposition 3. The fungi-free equilibrium $E_2 = \left(\frac{r_A - a_A}{b_A}, 0, \frac{q_O b_A + g(r_A - a_A)}{b_A a_O}\right)$ is feasible if

$$r_A > a_A \tag{7}$$

and it is stable if

$$\frac{hO_2}{k + k_O O_2} < a_F + cA_2 \tag{8}$$

hold.

Proof. If $F = 0$ in the system (2) we get

$$\begin{cases} r_A - a_A - b_A A = 0 \\ F = 0 \\ q_O + gA - a_O O = 0. \end{cases} \tag{9}$$

From the first equation of (9) it follows

$$A = \frac{r_A - a_A}{b_A}.$$

Thus for the nonnegativity of the algae population, (7) must hold. From the third equation instead we get the equilibrium value of the oxygen.

$$O = \frac{q_O b_A + g(r_A - a_A)}{b_A a_O}.$$

The characteristic polynomial associated to E_1 is once again easily obtained,

$$\det(J-\mu) = (-r_A + a_A - \mu)(-a_O - \mu)\left(\frac{hO_2}{k+k_O O_2} - a_F - cA_2 - \mu\right) = 0,$$

as well as its eigenvalues

$$\begin{aligned}
\mu_1 &= -r_a + a_A < 0 \\
\mu_2 &= -a_O < 0 \\
\mu_3 &= \frac{hO_2}{k+k_O O_2} - a_F - cA_2
\end{aligned}$$

Requiring $\mu_3 < 0$ we get (8). □

In Figure 2 we show that for a chosen set of parameters the stability of the fungi-free equilibrium, E_2, is attained.

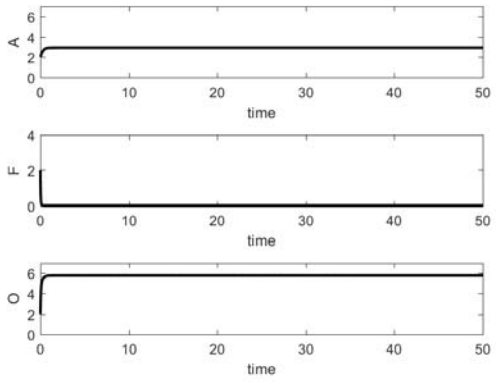


Figure 2: The equilibrium E_2 is stable for the parameter values $r_A = 8.90096$, $a_A = 4.14886$, $b_A = 1.62848$, $c = 0.0691916$, $h = 11.8402$, $k = 1.92602$, $k_O = 16.7554$, $a_F = 18.7713$, $b_F = 15.4976$, $q_O = 69.5303$, $g = 13.847$, $a_O = 19.037$, $f = 1$.

Proposition 4. The algae-free point is in fact a set of multiple equilibria, namely $(0, F_3, O_3)$, $(0, F_4, O_4)$ and $(0, F_5, O_5)$. Of them, only one is feasible, if

$$O > \frac{a_F k}{h - a_F k_O}, \tag{10}$$

and it is stable if

$$r_A < a_A + cF_3. \tag{11}$$

 ISBN: 978-84-608-6082-2

Proof. **Part 1: existence**

For $A = 0$ the system (2) becomes

$$\begin{cases} A = 0 \\ \dfrac{hO}{k + k_O O} - a_F - b_F F = 0 \\ q_O - a_O O - f\dfrac{hOF}{k + k_O O} = 0. \end{cases} \tag{12}$$

Solving the second equation with respect to F we get

$$F = \frac{O(h - a_F k_O) - a_F k}{b_F(k + k_O O)}. \tag{13}$$

Condition (10) arises by requiring the positivity of the expression (13). Note that the opposite case obtained when $h - a_F k_O < 0$, cannot arise,

$$O < \frac{a_F k}{h - a_F k_O},$$

because from it, $O < 0$ follows, which is impossible.

Substituting the expression (13) for F into the third equation of the system (12) we obtain the following third degree equation in O

$$aO^3 + bO^2 + cO + d = 0, \tag{14}$$

with

$$\begin{aligned} a &= -a_O b_F k_O^2 < 0 \\ b &= q_O b_F k_O^2 - 2a_O b_F k k_O - fh^2 + f h a_F k_O \\ c &= 2q_O b_F k k_O - a_O b_F k^2 + f h a_F k \\ d &= q_O b_F k^2 > 0. \end{aligned}$$

Since $a < 0$ and $d > 0$ by Descartes' rule of signs the third degree polynomial (14) in O has at least one positive root. We are able to show that there is exactly one such root. In the next Table the only four possible cases are summarized.

Cases	a	b	c	d	number of real positive roots
1)	−	+	−	+	3 (impossible)
2)	−	+	+	+	1
3)	−	−	−	+	1
4)	−	−	+	+	1

The first case is impossible, in fact assuming that $b > 0$ and $c < 0$ we find

$$a_O b_F k + \frac{fh^2}{k_O} < q_O b_F k_O - a_O b_F k + fha_F < -q_O b_F k_O.$$

But this is a contradiction, because the term in the middle should be less than a negative term (on the right) and greater than a positive one (on the left side).

Thus there is only one positive equilibrium

$$E_3 = \left(0, \frac{O_3(h - a_F k_O) - a_F k}{b_F(k + k_O O_3)}, O_3\right) \quad \text{if} \quad O_3 > \frac{a_F k}{h - a_F k_O},$$

with O_3 the real positive root of (14).

Part 2: stability

To study the stability of the equilibrium point we evaluate the Jacobian matrix (3) at E_3. The resulting characteristic polynomial is

$$\begin{aligned}\det(J - \mu) &= (r_A - a_A - cF_3 - \mu)\left\{\mu^2 + \left(a_F + 2b_F F_3 + a_O + \frac{fhkF_3}{(k + k_O O_3)^2} + \right.\right. \\ &\left.\left. - \frac{hO_3}{k + k_O O_3}\right)\mu + \left(a_O + \frac{fhkF_3}{(k + k_O O_3)^2}\right)(a_F + 2b_F F_3) - \frac{a_O h O_3}{k + k_O O_3}\right\} = 0.\end{aligned}$$

The eigenvalue $\mu_1 = r_A - a_A - cF_3$ is negative if (11) holds, while the roots of the quadratic polynomial in μ are negative with no further conditions. It turns out that both coefficients of the terms of the two lowest degrees in μ are positive. In fact, substituting F_3, (13), for the coefficient of μ we get

$$\begin{aligned}&\left(a_F + \frac{2hO_3}{k + k_O O_3} - \frac{2(O_3 a_F k_O + a_F k)}{k + k_O O_3} + a_O + \frac{fhkF_3}{(k + k_O O_3)^2} - \frac{hO_3}{k + k_O O_3}\right) \\ &= \left(a_F + \frac{hO_3}{k + k_O O_3} - 2a_F + a_O + \frac{fhkF_3}{(k + k_O O_3)^2}\right) \\ &= \frac{O_3(h - a_F k_O) - a_F k}{b_F(k + k_O O_3)} + a_O + \frac{fhkF_3}{(k + k_O O_3)^2} = F_3 + a_O + \frac{fhkF_3}{(k + k_O O_3)^2} > 0.\end{aligned}$$

Similarly, for the constant term, by dividing by a_O, denoting by H is a positive term, we have:

$$a_F + 2b_F F_3 - \frac{hO_3}{k + k_O O_3} + H > 0.$$

□

Figure 3 shows that for a chosen set of parameters the algae-free equilibrium, E_3, is stably achieved.

 ISBN: 978-84-608-6082-2

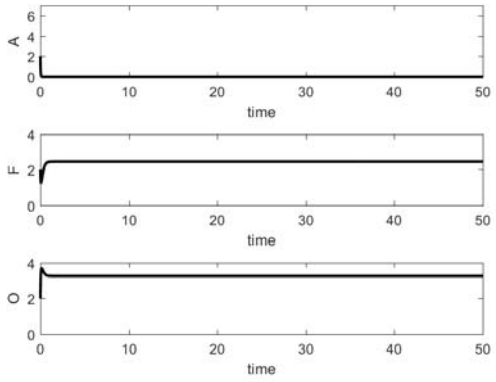


Figure 3: The equilibrium E_3 is stabile for the parameters $r_A = 4.85571$, $a_A = 13.2729$, $b_A = 11.6077$, $c = 12.7024$, $h = 11.4803$, $k = 2.3815$, $k_O = 1.11246$, $a_F = 1.40551$, $b_F = 1.95987$, $q_O = 65.3973$, $g = 16.7907$, $a_O = 15.374$, $f = 1$.

For the coexistence equilibrium point we have the following result

Proposition 5 There exists at least one feasible coexistence equilibrium $E_4 = (A^*, F^*, O^*)$ if the following three conditions hold:

$$r_A - a_A - b_A A^* > 0, \quad b_F b_A - c^2 > 0, \quad (a_F c + b_F)(k + k_O O^*)(r_A - a_A) > chO^* \tag{15}$$

and whenever it exists, it is stable.

Proof. To find the conditions for the existence of the coexistence equilibrium point from the first equation of the system (2) we get

$$F = \frac{r_A - a_A - b_A A}{c}$$

and substitute it into the remaining two equations. We solve these two equations with respect to A and we match the resulting expressions

$$A = \frac{(a_F c + b_F)(r_A - a_A)(k + k_O O) - chO}{(k + k_O O)(b_F b_A - c^2)} = \frac{fhO(r_A - a_A) + c(a_O O - q_O)(k + k_O O)}{cg(k + k_O O) + fhb_A O}.$$

Thus, we now have the following cubic polynomial in O:

$$a_1 O^3 + b_1 O^2 + c_1 O + d_1 = 0, \tag{16}$$

with

$$\begin{aligned}
a_1 &= -a_O k_O^2(b_F b_A - c^2) < 0 \quad \text{for} \quad b_F b_A - c^2 > 0\\
b_1 &= k_O(fhc + b_F k_O g)(r_A - a_A) + k_O(k_O q_O - 2a_O k)(b_F b_A - c^2) +\\
&\quad +(a_F k_O - h)(fhb_A + ck_O g)\\
c_1 &= k(fhc + 2b_F k_O g)(r_A - a_A) + k(2k_O q_O - a_O k)(b_F b_A - c^2) +\\
&\quad +chg(2a_F k_O - h) + a_F khb_A\\
d_1 &= a_F ck^2 g + k^2 q_O(b_F b_A - c^2) + b_F k^2 g(r_a - a_A) > 0.
\end{aligned}$$

Since for $b_F b_A - c^2 > 0$, $a_1 < 0$ and $d_1 > 0$ the polynomial (16) has at least one positive root O^* by the Descartes' rule of signs. For the feasibility of the equilibrium we need to have $F^* > 0$ and $A^* > 0$, providing thus the first and the third conditions in (8).

To study the stability, the characteristic polynomial of (3) evaluated at E_4 gives the cubic equation

$$\det(J - \mu) = \mu^3 + R\mu^2 + S\mu + P = 0$$

with

$$\begin{aligned}
R &= \left(b_A A^* + a_O + b_F F^* + \frac{fhkF^*}{(k + k_O O^*)^2}\right) > 0\\
S &= \left(A^* F^*(b_A b_F - c^2) + b_F a_O F^* + \frac{fhkF^*(b_A A^* + b_F F^*)}{(k + k_O O^*)^2} + \frac{h^2 kF^* O^*}{(k + k_O O^*)^3}\right) > 0\\
P &= \left((b_F b_A - c^2)\left(a_O A^* F^* + \frac{fhkF^{*2} A^*}{(k + k_O O^*)^2}\right) + \frac{b_A A^* h^2 kF^* O^*}{(k + k_O O^*)^3} + \frac{cghkA^* F^*}{(k + k_O O^*)^2}\right) > 0.
\end{aligned}$$

For $b_F b_A - c^2 > 0$, which holds by feasibility, the three eigenvalues are negative. Thus E_4 is stable whenever it is feasible. □

In Figure 4 we show that the coexistence equilibrium is stable for a selected set of parameter values.

In the next Table we summarize the feasibility conditions for the five equilibrium points of model (1).

In Figure 5 for a chosen set of parameters and the same initial conditions changing the value of q_O, the constant input of DO in the system, we obtain the stability of the coexistence equilibrium, E_4 on the left and of E_2 on the right. Thus, starting from the coexistence equilibrium, by decreasing the rate q_O at which oxygen is supplied into the system, we can obtain the fungi-free equilibrium.

 ISBN: 978-84-608-6082-2

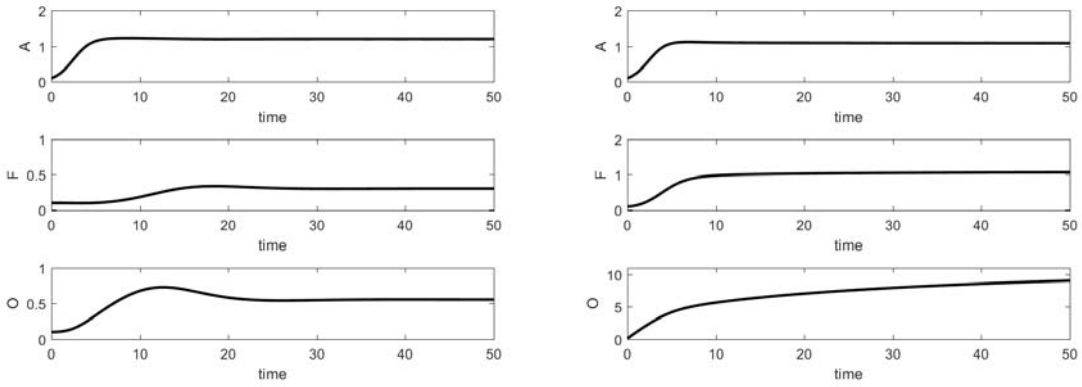


Figure 4: Two possible configurations for the equilibrium E_4, $q_O = 0$ and $q_O = 1$. It is stable in the following cases. Left: $q_O = 0$, $(1.20, 0.30, 0.55)$ is achieved for the parameters $r_A = 1$, $a_A = 0.001$, $b_A = 0.8$, $c = 0.1224$, $h = 1.1$, $k = 1$, $k_O = 1$, $a_F = 0.001$, $b_F = 0.8$, $q_O = 0$, $g = .1$, $a_O = 0.001$, $f = 1$. Right: $q_O = 1$, $(1.08, 1.08, 10.15)$ is obtained for the parameters $r_A = 1$, $a_A = 0.001$, $b_A = 0.8$, $c = 0.1224$, $h = 1.1$, $k = 1$, $k_O = 1$, $a_F = 0.001$, $b_F = 0.8$, $q_O = 1$, $g = .1$, $a_O = 0.001$, $f = 1$.

Eq.	Feasibility conditions	Stability conditions
E_0	$q_0 = 0$	$r_A < a_A$
E_1	none	$r_A < a_A$ and $\dfrac{hq_O}{ka_O + k_O q_O} < a_F$
E_2	$r_A > a_A$	$\dfrac{hO_2}{ka_O + k_O O_2} < a_F + cA_2$
E_3	$O_3 > \dfrac{a_F k}{h - a_F k_O}$	$r_A < a_A + cF_3$
E_4	$r_A - a_A - b_A A^* > 0, \quad b_F b_A - c^2 > 0$ $(a_F c + b_F)(k + k_O O^*)(r_A - a_A) > chO^*$	none

4 Conclusions and future work

A three dimensional, nonlinear mathematical model has been introduced and analysed. In addition to the trivial equilibrium, four additional equilibrium points have been found. Their stability was been completely analysed. For a chosen set of parameters with the same initial conditions we get the stability of the coexistence equilibrium, E_4, both in the absence, $q_O = 0$, and with full, $q_O = 1$, external oxygen supply, Figure 4. Thus the constant input of DO is not necessarily needed if the parameters are chosen appropriately, to have a viable

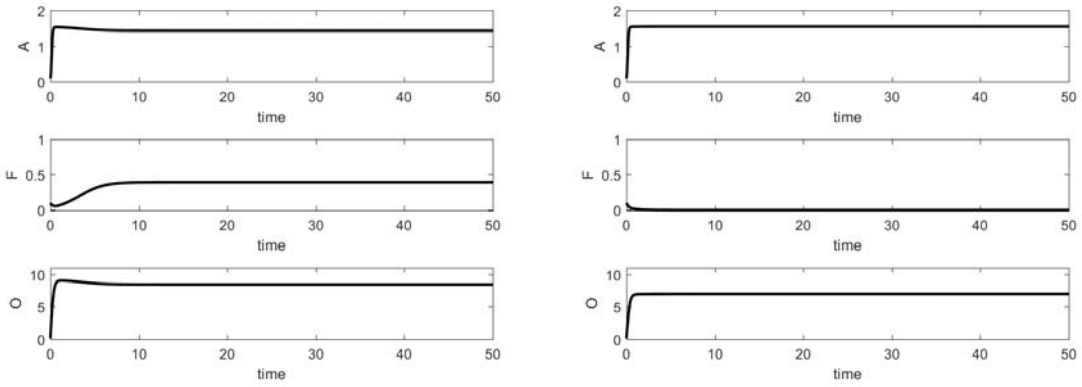


Figure 5: Left: the equilibrium E_4 is stable for $q_O = 30$; $r_A = 19.6445$, $a_A = 1.73234$, $b_A = 11.5828$, $c = 3.34324$, $h = 16.676$, $k = 12.4305$, $k_O = 0.517493$, $a_F = 3.04963$, $b_F = 1.3597$, $q_O = 30$, $g = 11.5323$, $f = 0.835718$, $a_O = 5.42261$ Right: the equilibrium E_2 at the stability $q_O = 20$ for $r_A = 19.6445$, $a_A = 1.73234$, $b_A = 11.5828$, $c = 3.34324$, $h = 16.676$, $k = 12.4305$, $k_O = 0.517493$, $a_F = 3.04963$, $b_F = 1.3597$, $q_O = 20$, $g = 11.5323$, $f = 0.835718$, $a_O = 5.42261$.

system. In fact algae contribution of DO to the system is enough for the fungi utilization. The simulations of Figure 5 instead show that the DO concentration should not drop below a critical threshold, because in such situation the fungi may disappear. Such a loss would be detrimental for the ecosystem.

One of the hypothesis of the model is the competition for food between algae and fungi, but in an indirect way the results indicate that algae help the fungi growth by producing DO.

In our future research we will compare the model introduced here (1) with another one in which the nutrient equation is also considered, as follows:

$$\begin{aligned}
\frac{dA}{dt} &= \frac{h_A N A}{k_A + k_N N} - e_A A - m_A A^2 \\
\frac{dF}{dt} &= k(N,O)F - e_F F - m_F F^2 \\
\frac{dN}{dt} &= q_N - r\frac{h_A N A}{k_A + k_N N} - sk(N,O)F - e_N N \\
\frac{dO}{dt} &= q_O + g_A A - e_O O - c_O k(N,O)F
\end{aligned} \tag{17}$$

with

$$k(N,O) = \frac{k_1 NO}{k_2 + k_3 N + k_4 O + k_3 k_4 NO}.$$

 ISBN: 978-84-608-6082-2

Acknowledgments

This work has been partially supported by the projects "Metodi numerici in teoria delle popolazioni" and "Metodi numerici nelle scienze applicate" of the Dipartimento di Matematica "Giuseppe Peano" of the Università di Torino.

References

[1] A. Anastasi, F. Spina, A. Romagnolo, V. Tigini, V. Prigione, G.C. Varese, *Integrated fungal biomass and activated sludge treatment for textile wastewaters bioremediation*, Bioresour Technol. **123:** 106-111 (2012).

[2] I.M. Bulai, E. Venturino, *Biodegradation of organic pollutants in a water body*, Journal of Mathematical Chemistry 1-17 (2016).

[3] A. Goyal, R. Sanghi, A.K. Misra, J.B. Shukla, *Modeling and analysis of the removal of an organic pollutant from a water body using fungi*, Appl. Math. Model. 38 (2014) 4863–4871.

[4] A.K. Misra, *Modeling the depletion of dissolved oxygen in a lake due to submerged macrophytes*, Nonlinear Analysis: Modelling and Control 15(2) (2010) 185–198.

[5] Han Li Qiao, E. Venturino, *A model for an aquatic ecosystem*, ICNAAM 2015, to appear in AIP Conference Proceedings.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Mathematical Aspects on Traffic of Incompressible Worms on Simple Circular Structures

Alexander P. Buslaev[1] **and Marina V. Yashina** [2]

[1] *Department of Mathematics, Moscow Automobile and Road State Tech. Univ. (MADI) and MTUCI*

[2] *Department of Mathematical Cybernetics and IT, Moscow Tech. Univ. of Communications and Informatics (MTUCI) and MADI*

emails: apal2006@yandex.ru, yash-marina@yandex.ru

Abstract

The dynamical system consisted of a set of clusters (arc segments) uniformly moving on circles is considered. Circles are connected with each other at common points named nodes. Clusters can not pass the common node at the same time. This situation is prohibited. If there is a competition then a rule of FIFO types is performed.

We discuss the average characteristics of the system and other quality properties. The problem is the formalization of a broad class of transportation problems.

Key words: Dynamical system, cluster model, transport-logistic problem, regular networks, synegry state, collapse state
MSC 2000: AMS codes 90B20, 90B10

1. Introduction

We consider N similar circuits - circles with standard angle system of coordinates $\varphi \in [0, 1]$, Fig. 1.

On each circle, Fig. 2, there is located a worm $[a_i,\ b_i]$ with length $x_i = b_i - a_i$ in angle measure. The worm crawls counterclockwise in free state with velocity 1.

We denote by ε_{ij} the coordinate of point on i-th circuit, if it is *common* with the point ε_{ji} on j-th circuit. Then the pair $(\varepsilon_{ij}, \varepsilon_{ji})$ is a node.

For simplicity, we assume that circuits i and j are not more than one node. Suppose $\varepsilon_{ij} = \varepsilon_{ji} = -1$, if $i = j$ or (i, j)-node does not exist because i-th circuits and j-th circuits not have a common point.

 ISBN: 978-84-608-6082-2

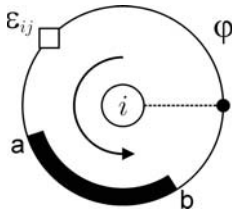


Figure 1: A worm (cluster) on a circle with standard angle system of coordinates

2. Dynamics

The basic rule of worms movement is that at the same time two worms $[a_i,\ b_i]$ and $[a_j,\ b_j]$ can not cover a common node, i.e.

$$(\varepsilon_{ij} \in [a_i,\ b_i]) \cap (\varepsilon_{ji} \in [a_j,\ b_j]) \neq \emptyset.$$

This rule of behavior is named FIFO in computer sciences.

If it will be a conflict at arrival to the node, then decision can be found using of coin tossing. The worm coming to not vacant node at velocity 1, stops and without changing its state expects when the node will be vacant. (Further, more complex alternative is compression process, [1]. It seems, the case is available for earthworms).

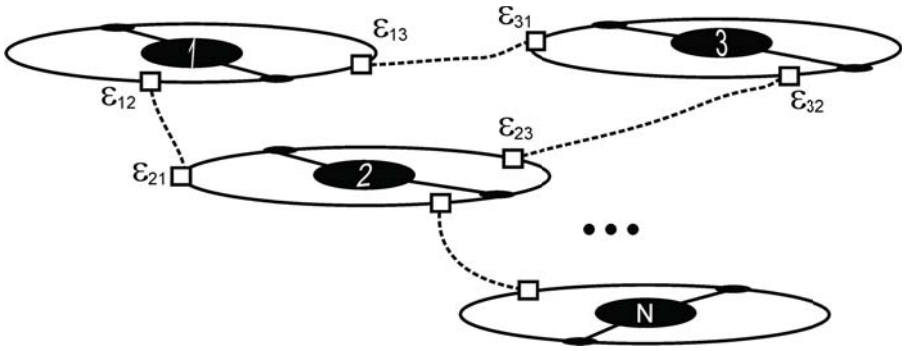


Figure 2: A set of N similar circuits, squares are common nodes

3. Problem formalization

The problem is to study the system behavior according to the parameters. Previously it was noted, [2], that such characteristic, as the average velocity, has features of spectral structure.

In particular, when worms have small size then the system demonstrates the property of self-organization (synergy), i.e. conflicts are disappearing, [2].

For exactlyness, for any admissible initial conditions after some finite time the system configuration becomes without conflicts and all clusters move with constant velocities.

So, *Markov process is reduced to determinate system.*

4. Short history of the problem

Cluster definition was introduced in [3] as limit state of connected flow in *follow-the-leader model* when *the leader is moving uniformly.*

Regular geometrical structure consisting from similar circuits, *necklace, honeycomb and chainmale,* [4], [5], is kindly inserted to considered formalization. Exact result for network of necklace type and a lot of numerical results for another cases afford ground for assumption that a threshold of synergy is accordingly 0.5, 1/3, 0.25. We need to note that final exact mathematical result is not obtained yet. Also it is obviously, that mathematical language is not found for similar kind of problems.

Let matrix $\varepsilon = (\varepsilon_{ij})$ consists from zeroes except the diagonal where - 1 numbers are located. This structure is named *traffic flower* and is developed on the work of Buslaev A.P and Tatashev A.G. in [7]. The main result is that *if total load, i.e. total length of worms (non necessarily identical) is less them length of the contour, than system becomes to without conflict state from any admissible conditions.*

5. The same individuals in the regular supporter (Circulant)

We assume (i, j)-node is simple with degree 1 (multiplicity), if each common point does not belong to more than two contours.

Then function

$$\varepsilon_{ij} = \varepsilon(i, j)$$

is one-to-one correspondence function for each argument.

Let us consider a system with the following matrix (1)

$$\varepsilon = \begin{pmatrix} -1 & \frac{1}{N-1} & \frac{2}{N-1} & \dots & \frac{N-1}{N-1} \\ \frac{N-1}{N-1} & -1 & \frac{1}{N-1} & \dots & \frac{N-2}{N-1} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{1}{N-1} & \frac{2}{N-1} & \dots & \frac{N-1}{N-1} & -1 \end{pmatrix} \tag{1}$$

and crawling objects with similar length l, $[a_i,\ b_i]$, $l \equiv b_i - a_i$, $i = 1, \dots, N$. Matrix (1) is a circulant formed by the vector with different coordinates. Therefore all nodes have simple

 ISBN: 978-84-608-6082-2

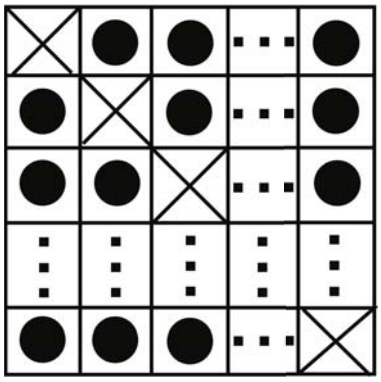


Figure 3: Square cell field $N \times N$

multiplicity. Geometric interpretation of GCM (Generalized Cluster Model) is as following. we have a square cell field $N \times N$, Fig. 3.

Worms with length l are cyclically crawling from left to right on all rows. With that, any pair of worms can not simultaneously crawl on detectors being symmetrically located in respect of the main diagonal. Cells with crosses denote that in those cells the space does not exist. Hence the distance between adjacent detectors in horizontal direction is equal to $(N-1)^{-1}$. Problem is to describe the system behavior.

6. Mathematics of incompressible worms on circulant

(6.1) *The set of admissible initial configurations is not empty if and only if* $l \leq 0.5$.

Proof.

$$[a_i(0),\ b_i(0)] = \left[\frac{i-1}{N-1}, l + \frac{i-1}{N-1}\right], \tag{2}$$

$i = 1, \ldots, N$. Then system (2) is correct when $l \leq 0.5$.

If $l > 0.5$, then for any location of worms there exists a conflict, since all worms cover more than half detectors.

Hence, there are at least a pair (i, j) covering symmetrical detectors.

(6.2) Let $f_i(t, x)$ be a characteristic function of i-th worm, $1 \leq i \leq N$,

$$f_i(t,x) = \{1,\ \text{if } x \in [a_i(t),\ b_i(t)]\,; 0,\ \text{if } x \notin [a_i(t),\ b_i(t)]\} \tag{3}$$

Denote by

$$g_{ij}(t,x) = f_i(t,x) f_j\left(t + \frac{j-i}{N-1}, x\right), \tag{4}$$

$$\Theta_{ij}(t) = \text{mes}\,\{x, g_{ij}(t,x) = 1\}\,. \tag{5}$$

Obviously, the value $\Theta ii(t)$ coincides with the delay time of one worm of worms pair, in condition that none of this pair was in conflict previously. And it is true $\Theta_{ij}(t) = \Theta_{ji}(t)$.

(6.3) Let

$$\Theta(t) = \sum_{1 \le i < j \le N} \Theta_{ij}(t).$$

If at the time t_0 $\Theta(t_0) = 0$, then $\Theta(t) \equiv 0$ at $t > t_0$ and the system is without conflicts.

(6.4) If $l < \frac{1}{N-1}$, then function $\Theta(t)$ decreases monotonically and within a finite time becomes to 0.

7. Some numerical results

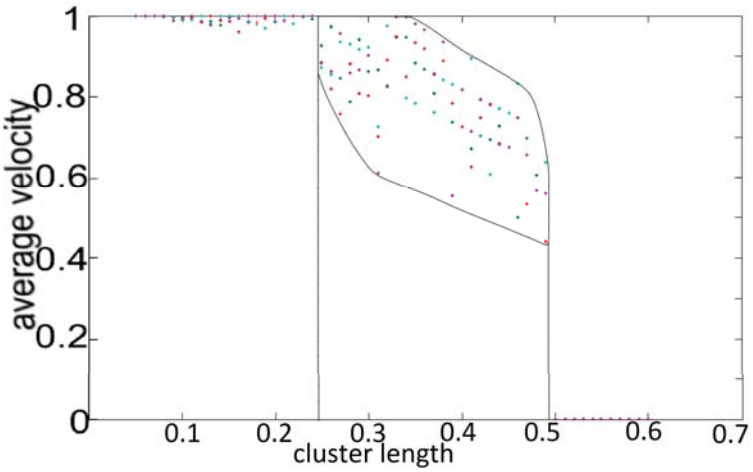


Figure 4: The dependence of average velocity on worm lengths on circulant network $N = 5$

Computer modeling of dynamical systems with such networks has been completed in the cases with the number of circuit is $N = 5$, Fig. 4.

Length of clusters – worms were assumed to be equal to all the circles, its value l were varied from 0 to 1 in increments 0.1.

The results have been showed that, for the network with 3 circles, the system velocity is equal a value closed to 1. Then, at the increasing of length to a threshold value $l = 0.5$, the system velocity decreases sharply to zero.

With the increase of the number of circles to 5 in the network, the threshold value of the system state changing is equal to the value closed to $l = 0.5$. At the same time we can observed that the system velocity decreases to zero more smooth by increasing of the cluster length.

 ISBN: 978-84-608-6082-2

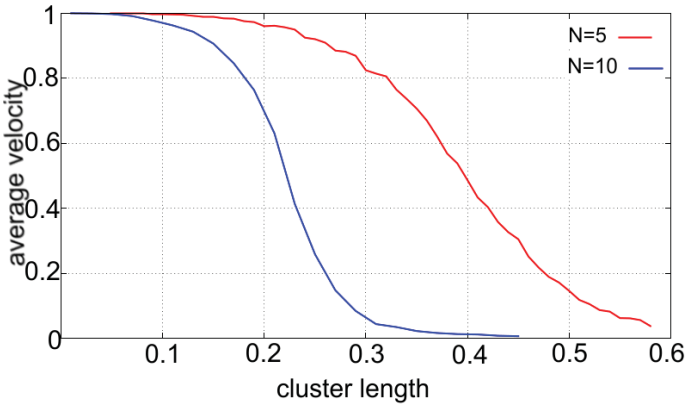


Figure 5: The comparison of circulant networks with $N = 5$ and $N = 10$

Further the increasing of circles number at the network is correspond to decreasing of the threshold value, Fig. 5.

8. Problems and Future Works

(8.1) it is important to study necessary and sufficient conditions of the correctness, the existence of consistent initial conditions on depending the system matrix, ε, and other parameters.

(8.2) Sufficient conditions of synergy and collapse states of the dynamical system.

(8.3) The existence and evaluation of state function.

The authors would like to thank the student Kuchelev D.A. (MTUCI) for provided computer-based fragments of numerical calculus.

References

[1] A. P. Buslaev , A. G. Tatashev, M. V. Yashina. *Cluster Flow Models and Properties of Appropriate Dynamic Systems.* Journal of Applied Functional Analysis. **8(1)** (2013) 54–76.

[2] Kozlov V. V., Buslaev A. P., Tatashev A. G., Yashina M. V. *On dynamical systems modelling for transport and communication.* Int. Conf. on Computational and Mathematical Methods in Science and Engineering (CMMSE 2014), Costa Ballena, Rota, Cadiz (Spain) July 3rd-7th, (2014).

[3] Bugaev A. S., Buslaev A. P., Kozlov V. V., Yashina M. V. *Distributed Problems of Monitoring and Modern Approaches to Traffic Modeling.* 14th International IEEE Conference on Intelligent Transportation Systems (ITSC 2011), Washington, USA, 5-7.10.2011. (2011) 477 – 481.

[4] Kozlov V. V., Buslaev A. P., Tatashev A. G. *Monotonic walks on a necklace and a coloured dynamic vector.* International Journal of Computer Mathematics, Taylor & Francis. **92:9** (2015), 1910-1920.

[5] Kozlov V. V., Buslaev A. P., Tatashev A. G., Yashina M. V. *Monotonic walk of particles on chainmail and colored matrices* Int. Conf. on Computational and Mathematical Methods in Science and Engineering (CMMSE 2014). Costa Ballena, Rota, Cadiz, Spain, 3-7 July, (2014).

[6] Buslaev A. P., Strusinskiy P. M. *On qualitative properties of incompressible cluster flow model on the ring network* AASRI Conf. on Sports Engineering and Computer Science (SECS 2014), AASRI Procedia. Conf. on Circuit and Signal Processing (CSP 2014), Elsevier, **9** (2014) 114-122.

[7] Buslaev A. P., Tatashev A. G. *Bernoulli Algebra on Common Fractions and Generalized Oscillations.* Journal of Mathematics Research. **8(3)** (2016) (to appear)

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

QL-fuzzy Implications by Means of Overlap and Grouping Functions

H. Bustince[1], M. Elkano[1], G. Dimuro[2], B. Bedregal[3], M. Sesma-Sara[1] and G. Lucca[1]

[1] *Department of Automation and Computing, Universidad Publica de Navarra*

[2] *Centro de Ciência Computacionais, Universidade Federal do Rio Grande*

[3] *Departamento de Informática e Matemática Aplicada, Universidade Federal do Rio Grande do Norte*

emails: bustince@unavarra.es, mikel.elkano@unavarra.es, gracaliz@gmail.com, bedregal@dimap.ufrn.br, mikel.sesma@unavarra.es, lucca.112793@e.unavarra.es

Abstract

In this work we discuss the construction of a particular case of QL-implications using overlap and grouping functions. *Key words: overlap function, grouping function, QL implication*

1 Introduction

Fuzzy implications, which generalize the classical implication to the fuzzy setting, are a very relevant tool for many applications [1, 2]. Recall that a fuzzy implication is just a mapping $I : [0,1]^2 \to [0,1]$ such that $I(0,0) = I(0,1) = I(1,1) = 1$, $I(1,0) = 0$, I is decreasing in the first variable and increasing in the second variable. The boundary conditions required in this definition make that fuzzy implications are a generalization to the fuzzy setting of the classical concept of implication.

The different approaches to the concept of implication in the classical setting lead to different operators when they are generalized for fuzzy sets. including the so-called R-implications, S-implications and QL-implications. The latter, in which this paper is focused on, extend the implication

$$p \to q \equiv \neg p \vee (p \wedge q)$$

 ISBN: 978-84-608-6082-2

which is used in quantum logic. The usual way in which this implication is extended is by considering a t-norm T, a t-conorm S and a negation N to replace $\wedge$, $\vee$ and $\neg$, respectively [3]. The choice of these functions, and specially the t-norm and the t-conorm, may be too restrictive, since, some demanding properties such as associativity are required in order to define them. Associativity, however, does not play any role in the generalization [4, 5, 6]. For this reason, in this contribution we analyze the replacement of the t-norm by an overlap function and the t-conorm by a grouping function [5, 6], which have attracted a lot of attention in recent years [7, 8, 9, 10]. This replacement has already led to very interesting results in some applications of t-norms or t-conorms where associativity was not necessary (for instance, in classification [11, 12], in image processing [13] and in decision making [14, 15, 16].

The structure of this paper is as follows.

2 Preliminaries

In this section we present some preliminary concepts that are needed for the rest of the contribution.

Definition 2.1 *A function $N : [0,1]^2 \to [0,1]$ is said to be a fuzzy negation if the following conditions hold:*

(N1) *N satisfies the Boundary Conditions: $N(0) = 1$ and $N(1) = 0$;*

(N2) *N is decreasing: if $x \leq y$ then $N(y) \leq N(x)$.*

Example 2.1 *The least fuzzy negation $N_\perp$ and the greatest fuzzy negation $N_\top$ are defined, respectively, by*

$$N_\perp(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \in]0,1] \end{cases} \tag{1}$$

and

$$N_\top(x) = \begin{cases} 0 & \text{if } x = 1 \\ 1 & \text{if } x \in [0,1[, \end{cases} \tag{2}$$

which are both crisp. In fact, one has that $N_\perp = N_0$ and $N_\top = N^1$. The standard negation or Zadeh negation is given by $N_Z(x) = 1 - x$.

A fuzzy negation is called frontier if it satisfies the property:

(N3) *$N(x) \in \{0,1\}$ if and only if $x = 0$ or $x = 1$.*

A fuzzy negation is called crisp if it satisfies the property:

(N4) *For all $x \in [0,1]$, $N(x) \in \{0,1\}$.*

A fuzzy negation is called non-filling if it satisfies the property:

(N5) $N(x) = 1$ *if and only if* $x = 0$.

Definition 2.2 *[17] A function* $A : [0,1]^n \to [0,1]$ *is said to be an n-ary aggregation operator if the following conditions hold:*

(A1) *A is increasing*[1] *in each argument: for each* $i \in \{1, \ldots, n\}$, *if* $x_i \leq y$, *then* $A(x_1, \ldots, x_n) \leq A(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$;

(A2) *A satisfies the Boundary conditions:* $A(0, \ldots, 0) = 0$ *and* $A(1, \ldots, 1) = 1$.

The main classes of aggregation we are going to consider in this work are those of overlap and grouping functions[7, 5, 6, 8, 18, 19, 9, 20, 10, 21, 13].

Definition 2.3 *A bivariate function* $O : [0,1]^2 \to [0,1]$ *is said to be an overlap function if it satisfies the following conditions:*

(O1) O *is commutative;*

(O2) $O(x, y) = 0$ *if and only if* $x = 0$ *ou* $y = 0$;

(O3) $O(x, y) = 1$ *if and only if* $x = y = 1$;

(O4) O *is increasing;*

(O5) O *is continuous.*

Definition 2.4 *A bivariate function* $G : [0,1]^2 \to [0,1]$ *is said to be a grouping function if it satisfies the following conditions:*

(G1) G *is commutative;*

(G2) $G(x, y) = 0$ *if and only if* $x = y = 0$;

(G3) $G(x, y) = 1$ *if and only if* $x = 1$ *or* $y = 1$;

(G4) G *is increasing;*

(G5) G *is continuous.*

Definition 2.5 *A function* $I : [0,1]^2 \to [0,1]$ *is a fuzzy implication if, for each* $x, y, z \in [0,1]$, *it holds that:*

[1] In this paper, a increasing (decreasing) function does not need to be strictly increasing (decreasing).

(I1) *First place antitonicity: if* $x \leq y$ *then* $I(y,z) \leq I(x,z)$*;*

(I2) *Second place isotonicity: if* $y \leq z$ *then* $I(x,y) \leq I(x,z)$*;*

(I3) *Boundary condition 1:* $I(0,0) = 1$*;*

(I4) *Boundary condition 2:* $I(1,1) = 1$*;*

(I5) *Boundary condition 3:* $I(1,0) = 0$*.*

From an analysis of the different properties demanded to implication functions see [3, 22].

3 QL-Implications Derived from Tuples (O, G, N)

In this section we present the main results in this contribution. In particular, we analyze the construction of an analog of QL-implications using overlap and grouping functions, as follows.

Definition 3.1 *A function* $I : [0,1]^2 \to [0,1]$ *is a QL-operation derived from a tuple* (O, G, N) *if there exist an overlap function* $O : [0,1]^2 \to [0,1]$*, a grouping function* $G : [0,1]^2 \to [0,1]$ *and a fuzzy negation* $N : [0,1] \to [0,1]$*, such that*

$$I(x,y) = G(N(x), O(x,y)), \tag{3}$$

for all $x, y \in [0,1]$*. We denote such QL-operation by* $I_{O,G,N}$*.*

This construction satisfies some of the properties demanded to QL-implications.

Proposition 3.1 *If* $I_{O,G,N}$ *is a QL-operation derived from a tuple* (O, G, N)*, then* $I_{O,G,N}$ *satisfies (I2), (I3), (I4) and (I5).*

Proof.
(I2) Consider $y \leq z$. Then, since G and O are increasing, it holds that:

$$I_{O,G,N}(x,y) = G(N(x), O(x,y)) \leq G(N(x), O(x,z)) = I_{O,G,N}(x,z).$$

(I3) It follows that $I_{O,G,N}(0,0) = G(N(0), O(0,0)) = G(1,0) = 1$.
(I4) It follows that $I_{O,G,N}(1,1) = G(N(1), O(1,1)) = G(0,1) = 1$.
(I5) It follows that $I_{O,G,N}(1,0) = G(N(1), O(1,0)) = G(0,0) = 0$.

Lemma 3.1 *Any QL-operation derived from the tuple* $(O, G, N_\top)$ *is of the form:*

$$I_{O,G,N_\top}(x,y) = \begin{cases} G(0, O(1,y)) & \text{if } x = 1 \\ 1 & \text{if } x < 1. \end{cases} \tag{4}$$

Proof. It follows that

$$\begin{aligned} I_{O,G,N_\top}(x,y) &= G(N_\top(x), O(x,y)) \\ &= \begin{cases} G(0, O(1,y)) & \text{if } x = 1 \\ G(1, O(x,y)) & \text{if } x < 1 \end{cases} \\ &= \begin{cases} G(0, O(1,y)) & \text{if } x = 1 \\ 1 & \text{if } x < 1 \end{cases} \end{aligned}$$

We can characterize when we recover a fuzzy implication in the following way.

Theorem 3.1 *A QL-operation derived from the tuple (O, G, N) is a fuzzy implication if and only if $N = N_\top$.*

Proof. ($\Rightarrow$) Suppose that $N \neq N_\top$. Then, there exists $x \in]0,1[$, such that $N(x) < 1$. It follows that:

$$I_{O,G,N}(x,1) = G(N(x), O(x,1)) < 1 = I_{O,G,N}(1,1),$$

and, thus, $I_{O,G,N}$ does not satisfy **(I1)**, which is a contradiction. Therefore, one concludes that $N = N_\top$.
($\Leftarrow$) By Lemma 3.1, one has that

$$I_{O,G,N_\top}(x,y) = \begin{cases} G(0, O(1,y)) & \text{if } x = 1 \\ 1 & \text{if } x < 1. \end{cases}$$

Then, if $x \leq z$ and $x = 1$, the result is immediate. Suppose that $x < 1$. It follows that $I_{O,G,N}(x,y) = 1 \geq I_{O,G,N}(z,y)$. Thus, $I_{O,G,N}$ satisfies **(I1)**, and the result follows from Proposition 3.1 **(i)**.

Proposition 3.2 *Let $O : [0,1]^2 \to [0,1]$ and $G : [0,1]^2 \to [0,1]$ be overlap and grouping functions, respectively, and $N_\top : [0,1] \to [0,1]$ the greatest fuzzy negation. The following statements hold:*

(i) If O has 1 as neutral element, then the QL-operation derived from the tuple $(O, G, N_\top)$ is a fuzzy implication of the form:

$$I_{O,G,N_\top}(x,y) = \begin{cases} 1 & \text{if } x < 1 \ \vee \ y = 1; \\ G(0,y) & \text{if } x = 1 \ \wedge \ y < 1. \end{cases} \tag{5}$$

(ii) If G has 0 as neutral element, then the QL-operation derived from the tuple $(O, G, N_\top)$ of the form:

$$I_{O,G,N_\top}(x,y) = \begin{cases} O(1,y) & \text{if } x = 1; \\ 1 & \text{if } x < 1. \end{cases} \tag{6}$$

is a fuzzy implication.

Proof. It follows that:

(i) It follows that:

(a) If $x < 1$, then

$$I_{O,G,N_\top}(x, y) = G(N_\top(x), O(x, y)) = G(1, O(x, y)) = 1.$$

(b) If $y = 1$, then

$$I_{O,G,N_\top}(x, 1) = G(N_\top(x), O(x, 1)) = G(N_\top(x), x) = 1.$$

(c) If $x = 1$ and $y < 1$, then $I_{O,G,N_\top}(1, y) = G(N_\top(1), O(1, y)) = G(0, y)$.

Thus $I_{O,G,N_\top}$ is a QL-operation and, by Theorem 3.1, it is immediate that $I_{O,G,N_\top}$ is a fuzzy implication.

(ii) It follows that:

(a) If $x = 1$, then:

$$I_{O,G,N_\top}(1, y) = G(N_\top(1), O(1, y)) = G(0, O(1, y)) = O(1, y).$$

(a) If $x < 1$, then:

$$I_{O,G,N_\top}(x, y) = G(N_\top(x), O(x, y)) = G(1, O(x, y)) = 1.$$

Thus since $I_{O,G,N_\top}$ is a QL-operation, then, by Theorem 3.1, it is immediate that $I_{O,G,N_\top}$ is a fuzzy implication.

4 Conclusions

In this contribution we have discussed the construction of analogs of QL fuzzy implications using overlap and grouping functions. We have characterized when we recover an implication functions.

In the future we intend to analyze the relation of fuzzy implications built in this way with other classes of implications.

Acknowledgements

This work has been partially supported by project TIN2013-40765-P of the Spanish Ministry of Science.

References

[1] M. Baczyński, On the applications of fuzzy implication functions, in: V. E. Balas, J. Fodor, A. R. Várkonyi-Kóczy, J. Dombi, L. C. Jain (Eds.), Soft Computing Applications, Vol. 195 of Advances in Intelligent Systems and Computing, Springer, Berlin, 2013, pp. 9–10.

[2] M. Baczyński, G. Beliakov, H. Bustince Sola, A. Pradera (Eds.), Advances in Fuzzy Implication Functions, Vol. 300 of Studies in Fuzziness and Soft Computing, Springer, Berlin, 2013.

[3] M. Baczyński, B. Jayaram, Fuzzy Implications, Springer, Berlin, 2008.

[4] H. Bustince, J. Fernández, R. Mesiar, J. Montero, R. Orduna, Overlap index, overlap functions and migrativity, in: Proceedings of IFSA/EUSFLAT Conference, 2009, pp. 300–305.

[5] H. Bustince, J. Fernandez, R. Mesiar, J. Montero, R. Orduna, Overlap functions, Nonlinear Analysis: Theory, Methods & Applications 72 (3-4) (2010) 1488–1499.

[6] H. Bustince, M. Pagola, R. Mesiar, E. Hüllermeier, F. Herrera, Grouping, overlaps, and generalized bientropic functions for fuzzy modeling of pairwise comparisons, IEEE Transactions on Fuzzy Systems 20 (3) (2012) 405–415.

[7] B. C. Bedregal, G. P. Dimuro, H. Bustince, E. Barrenechea, New results on overlap and grouping functions, Information Sciences 249 (2013) 148–170.

[8] G. P. Dimuro, B. Bedregal, Archimedean overlap functions: The ordinal sum and the cancellation, idempotency and limiting properties, Fuzzy Sets and Systems 252 (2014) 39 – 54.

[9] G. P. Dimuro, B. Bedregal, H. Bustince, M. J. Asiáin, R. Mesiar, On additive generators of overlap functions, Fuzzy Sets and Systems DOI: http://dx.doi.org/10.1016/j.fss.2015.02.008.

[10] G. P. Dimuro, B. Bedregal, H. Bustince, R. Mesiar, M. J. Asiain, On additive generators of grouping functions, in: A. Laurent, O. Strauss, B. Bouchon-Meunier, R. R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems, Vol. 444 of Communications in Computer and Information Science, Springer International Publishing, 2014, pp. 252–261.

[11] M. Elkano, M. Galar, J. Sanz, H. Bustince, Fuzzy rule-based classification systems for multi-class problems using binary decomposition strategies: On the influence of

n-dimensional overlap functions in the fuzzy reasoning method, Information Sciences 332 (2016) 94–114.

[12] G. Lucca, G. P. Dimuro, V. Mattos, B. Bedregal, H. Bustince, J. A. Sanz, A family of Choquet-based non-associative aggregation functions for application in fuzzy rule-based classification systems, in: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, Los Alamitos, 2015, pp. 1–8. doi:10.1109/FUZZ-IEEE.2015.7337911.

[13] A. Jurio, H. Bustince, M. Pagola, A. Pradera, R. Yager, Some properties of overlap and grouping functions and their application to image thresholding, Fuzzy Sets and Systems 229 (2013) 69 – 90.

[14] E. Barrenechea, J. Fernandez, M. Pagola, F. Chiclana, H. Bustince, Construction of interval-valued fuzzy preference relations from ignorance functions and fuzzy preference relations. Application to decision making, Knowledge-Based Systems 58 (2014) 33 – 44.

[15] H. Bustince, E. Barrenechea, T. Calvo, S. James, G. Beliakov, Consensus in multi-expert decision making problems using penalty functions defined over a cartesian product of lattices, Information Fusion 17 (2014) 56–64.

[16] M. Elkano, M. Galar, J. Sanz, A. Fernández, E. Barrenechea, F. Herrera, H. Bustince, Enhancing multi-class classification in FARC-HD fuzzy classifier: On the synergy between n-dimensional overlap functions and decomposition strategies, IEEE Transactions on Fuzzy Systems 23 (5) (2015) 1562–1580.

[17] G. Beliakov, A. Pradera, T. Calvo, Aggregation Functions: A Guide for Practitioners, Springer, Berlin, 2007.

[18] G. P. Dimuro, B. Bedregal, On residual implications derived from overlap functions, Information Sciences 312 (2015) 78 – 88.

[19] G. P. Dimuro, B. Bedregal, On the laws of contraposition for residual implications derived from overlap functions, in: Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, IEEE, Los Alamitos, 2015, pp. 1–7. doi:10.1109/FUZZ-IEEE.2015.7337867.

[20] G. P. Dimuro, B. Bedregal, H. Bustince, J. Fernandez, M. Pagola, Fuzzy implications derived from overlap and grouping functions and the law of o-conditionality (to appear).

[21] G. P. Dimuro, B. Bedregal, R. H. N. Santiago, On (G, N)-implications derived from grouping functions, Information Sciences 279 (2014) 1 – 17.

[22] H. Bustince, P. Burillo, F. Soria, Automorphism, negations and implication operators, Fuzzy Sets and Systems 134 (2) (2003) 209–229.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The natural embedding of fuzzy preposets and its residual mapping

Inma P. Cabrera[1], Pablo Cordero[1] and Manuel Ojeda-Aciego[1]

[1] *Departamento de Matemática Aplicada, Universidad de Málaga. Spain*

emails: ipcabrera@uma.es, pcordero@uma.es, aciego@uma.es

Abstract

We continue the study of adjunctions, also called isotone Galois connections, in the framework of fuzzy preordered structures, which generalize fuzzy preposets by considering underlying fuzzy equivalence relations. Specifically, given a subset $X \neq \varnothing$ of a fuzzy structure $\mathcal{B} = \langle B, \approx_B \rangle$ together with any fuzzy preorder ρ_X on $\langle X, \approx_B \rangle$, we study necessary and sufficient conditions guaranteeing the existence of a fuzzy preorder relation ρ_B and a residual mapping (namely a right adjoint) $h\colon B \to X$ to the natural embedding $i\colon X \hookrightarrow B$.

Key words: Galois connection, Adjunction, Preorder, Fuzzy sets

1 Introduction

Adjunctions (together with the closely-related Galois connections) can be found in different areas, and it is common to find papers dealing with them either from a practical or a theoretical point of view. In the literature, one can find numerous papers on theoretical developments on adjunctions [1, 2, 4, 9, 12] and also on applications thereof [7, 8, 13, 14, 15, 16, 17, 18].

In previous works [11, 10], the authors have studied the following problem: given a mapping $f\colon \mathbb{A} \to B$ from a (fuzzy) preorder $\mathbb{A}$ into an unstructured set B, characterize when it is possible to construct a suitable (fuzzy) preorder relation on B for which there exists a mapping $g\colon B \to \mathbb{A}$ such that the pair (f, g) constitutes an adjunction. We continue the study of adjunctions, in the framework of fuzzy preordered structures, which generalize fuzzy preposets by considering underlying fuzzy equivalence relations.

In [6] we focused on the case of a fuzzy ordering ρ_A on A and a surjective mapping $f\colon \langle A, \approx_A \rangle \to \langle B, \approx_B \rangle$ compatible with respect to the fuzzy equivalences $\approx_A$ and $\approx_B$,

 ISBN: 978-84-608-6082-2

which can be seen as a solution in the first part in the canonical decomposition of f. In this paper, we somehow complete the study of the decomposition by considering the natural embedding of the image in the codomain, which we formulate in abstract terms as follows: given a subset $X \neq \varnothing$ of a fuzzy structure $\mathcal{B} = \langle B, \approx_B \rangle$ together with any fuzzy preorder ρ_X on $\langle X, \approx_B \rangle$, we study necessary and sufficient conditions guaranteeing the existence of a fuzzy preorder relation ρ_B and a residual mapping (namely a right adjoint) $h\colon B \to X$ to the natural embedding $i\colon X \hookrightarrow B$.

2 Preliminaries

The most common underlying structure for considering fuzzy generalizations of Galois connections is that of a complete residuated lattice $\mathbb{L} = (L, \leq, \top, \bot, \otimes, \to)$. As usual, supremum and infimum will be denoted by $\vee$ and $\wedge$, respectively. An $\mathbb{L}$-*fuzzy set* X on a universe U is a mapping $X\colon U \to L$ from U to the membership values structure L where $X(u)$ denotes the degree to which u belongs to X. Given two $\mathbb{L}$-fuzzy sets X and Y, X is said to *be included in* Y, denoted as $X \subseteq Y$, if $X(u) \leq Y(u)$ for all $u \in U$.

A mapping $R\colon U \times U \to L$ is a (binary) $\mathbb{L}$-*fuzzy relation on* U. An $\mathbb{L}$-fuzzy relation R is said to be:

(i) *Reflexive* if $R(a, a) = \top$ for all $a \in U$.

(ii) $\otimes$-*Transitive* if $R(a, b) \otimes R(b, c) \leq R(a, c)$ for all $a, b, c \in U$.

(iii) *Symmetric* if $R(a, b) = R(b, a)$ for all $a, b \in U$.

(iv) *Antisymmetric* if $R(a, b) = R(b, a) = \top$ implies $a = b$, for all $a, b \in U$.

From now on, when no confusion arises, we will omit the prefixes "$\mathbb{L}$-" and "$\otimes$-".

Definition 1 *A* fuzzy preposet *is a pair* $\mathbb{A} = \langle A, \rho_A \rangle$ *in which* ρ_A *is a reflexive and transitive fuzzy relation on* A.

Definition 2 *A fuzzy relation* $\approx$ *on* A *is said to be a:*

- Fuzzy equivalence relation *if* $\approx$ *is a reflexive, symmetric, and transitive fuzzy relation on* A.
- Fuzzy equality relation *if* $\approx$ *is a fuzzy equivalence relation satisfying that* $\approx(a, b) = \top$ *implies* $a = b$, *for all* $a, b \in A$.

We will use the infix notation for a fuzzy equivalence relation, that is: for a fuzzy equivalence relation $\approx\colon A \times A \to L$, we write $a_1 \approx a_2$ to refer to $\approx(a_1, a_2)$.

Definition 3 *A* fuzzy structure $\mathcal{A} = \langle A, \approx_A \rangle$ *is a set* A *endowed with a fuzzy equivalence relation* $\approx_A$.

Definition 4 *A* morphism *between two fuzzy structures $\mathcal{A}$ and $\mathcal{B}$ is a mapping $f\colon A \to B$ such that for $a_1, a_2 \in A$ the following inequality holds $(a_1 \approx_A a_2) \leq (f(a_1) \approx_B f(a_2))$. In this case, we write $f\colon \mathcal{A} \to \mathcal{B}$, and we say that f is* compatible *with $\approx_A$ and $\approx_B$.*

It is worth to note that the composition of two morphisms is a morphism.

Definition 5 ([3]) *Let $\approx_A$ be a fuzzy equivalence relation on A. A fuzzy binary relation $\rho_A\colon A \times A \to L$ is said to be*

(i) $\approx_A$-reflexive *if $(a_1 \approx_A a_2) \leq \rho_A(a_1, a_2)$ for all $a_1, a_2 \in A$.*

(ii) $\otimes$-$\approx_A$-antisymmetric *if $\rho_A(a_1, a_2) \otimes \rho_A(a_2, a_1) \leq (a_1 \approx_A a_2)$ for all $a_1, a_2 \in A$.*

Definition 6 *Given a fuzzy structure $\mathcal{A} = \langle A, \approx_A \rangle$, the pair $\mathbb{A} = \langle \mathcal{A}, \rho_A \rangle$ will be called a* $\otimes$-$\approx_A$- fuzzy preordered structure *or simply* fuzzy preordered structure *(when there is no risk of confusion), if ρ_A is a fuzzy relation that is $\approx_A$-reflexive, $\otimes$-$\approx_A$-antisymmetric and $\otimes$-transitive.*

If the underlying fuzzy structure is not clear from the context, we will sometimes write a fuzzy preordered structure as a triplet $\mathbb{A} = \langle A, \approx_A, \rho_A \rangle$.

A reasonable approach to introduce the notion of fuzzy adjunction between fuzzy preordered structures $\mathbb{A}$ and $\mathbb{B}$ would be the following

Definition 7 *Let $\mathbb{A}$ and $\mathbb{B}$ be two fuzzy preordered structures. Given two morphisms $f\colon \mathcal{A} \to \mathcal{B}$ and $g\colon \mathcal{B} \to \mathcal{A}$, the pair (f, g) is said to be a* fuzzy adjunction *between $\mathbb{A}$ and $\mathbb{B}$ (briefly, $(f, g)\colon \mathbb{A} \leftrightharpoons \mathbb{B}$) if the following conditions hold:*

(G1) $(a_1 \approx_A a_2) \otimes \rho_A(a_2, g(b)) \leq \rho_B(f(a_1), b)$

(G2) $(b_1 \approx_B b_2) \otimes \rho_B(f(a), b_1) \leq \rho_A(a, g(b_2))$

for all $a, a_1, a_2 \in A$ and $b, b_1, b_2 \in B$.

It turns out that the previous definition is strongly related to the straightforward definition given in [5].

Theorem 1 *Let $\mathbb{A} = \langle \mathcal{A}, \rho_A \rangle$ and $\mathbb{B} = \langle \mathcal{B}, \rho_B \rangle$ be two fuzzy preordered structures, and consider two mappings $f\colon A \to B$ and $g\colon B \to A$. Then, the pair (f, g) is a fuzzy adjunction between $\mathbb{A}$ and $\mathbb{B}$ if and only if both mappings are morphisms and $\rho_A(a, g(b)) = \rho_B(f(a), b)$ for all $a \in A$ and $b \in B$.*

Corollary 1 *If a pair (f, g) is a fuzzy adjunction between two fuzzy preordered structures $\langle A, \approx_A, \rho_A \rangle$ and $\langle B, \approx_B, \rho_B \rangle$, then (f, g) is also an adjunction between the two fuzzy preposets $\langle A, \rho_A \rangle$ and $\langle B, \rho_B \rangle$.*

Conversely, if a pair (f, g) is an adjunction between two fuzzy preposets $\langle A, \rho_A \rangle$ and $\langle B, \rho_B \rangle$ then (f, g) is also a fuzzy adjunction between the two fuzzy preordered structures $\langle A, =, \rho_A \rangle$ and $\langle B, =, \rho_B \rangle$, where $=$ denotes the standard (crisp) equality.

 ISBN: 978-84-608-6082-2

3 The construction

The notion of contraction is introduced below:

Definition 8 *Let $\mathcal{B} = \langle B, \approx_B \rangle$ be a fuzzy structure, and a crisp subset $X \subseteq B$. A mapping $h\colon B \to X$ is said to be a* contraction *if it is a morphism $h\colon \mathcal{B} \to \langle X, \approx_B \rangle$ and $h(x) = x$ for all $x \in X$.*

Lemma 1 *Given a fuzzy structure $\mathcal{B} = \langle B, \approx_B \rangle$, a nonempty crisp subset $X \subseteq B$, a $\approx_B$-reflexive, $\otimes$-$\approx_B$-antisymmetric and $\otimes$-transitive fuzzy relation ρ_X on $\langle X, \approx_B \rangle$, and a contraction $h\colon B \to X$, the fuzzy relation $\mu_h\colon B \times B \to L$ defined below is reflexive*

$$\mu_h(b_1, b_2) = \begin{cases} \rho_X(b_1, h(b_2)) & \text{if } b_1 \in X\,, \\ b_1 \approx_B b_2 & \text{if } b_1 \notin X\,. \end{cases}$$

The fuzzy relation μ_h above will be called *h-reflexive closure of ρ_X*; the term 'closure' makes sense since any fuzzy relation ρ_B which extends $\mathcal{B} = \langle B, \approx_B \rangle$ to be a fuzzy preordered structure for which there exists a contraction h such that $(i, h)\colon \langle X, \approx_B, \rho_X \rangle \leftrightharpoons \langle B, \approx_B, \rho_B \rangle$ should satisfy $\mu_h \leq \rho_B$.

Although μ_h is reflexive, it might fail to be transitive, as shown in Example 1, therefore the transitive closure of μ_h, denoted μ_h^t, should be contained in ρ_B as well.

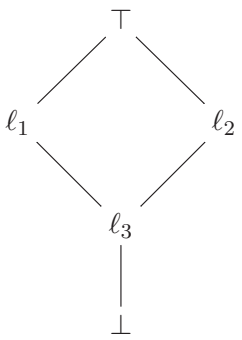


Figure 1: The lattice $(L, \leq)$

Example 1 *Consider the residuated lattice $\mathbb{L} = (L, \leq, \top, \bot, \otimes, \to)$ where $(L, \leq)$ is depicted in Figure 1 and the product $\otimes$ is the following one:*

$\otimes$	$\bot$	ℓ_1	ℓ_2	ℓ_3	$\top$
$\bot$	$\bot$	$\bot$	$\bot$	$\bot$	$\bot$
ℓ_1	$\bot$	ℓ_1	$\bot$	$\bot$	ℓ_1
ℓ_2	$\bot$	$\bot$	ℓ_2	$\bot$	ℓ_2
ℓ_3	$\bot$	$\bot$	$\bot$	$\bot$	ℓ_3
$\top$	$\bot$	ℓ_1	ℓ_2	ℓ_3	$\top$

Consider $B = \{x_1, x_2, b\}$, the subset $X = \{x_1, x_2\}$ and the two $\mathbb{L}$-fuzzy relations below:

$\approx_B$	x_1	x_2	b
x_1	$\top$	ℓ_2	ℓ_1
x_2	ℓ_2	$\top$	ℓ_2
b	ℓ_1	ℓ_2	$\top$

ρ_X	x_1	x_2
x_1	$\top$	ℓ_2
x_2	ℓ_2	$\top$

For the contraction $h\colon B \to X$, where $h(x_1) = h(b) = x_1$ and $h(x_2) = x_2$, the h-reflexive closure of ρ_X is given in the following table:

μ_h	x_1	x_2	b
x_1	$\top$	ℓ_2	$\top$
x_2	ℓ_2	$\top$	ℓ_2
b	ℓ_1	ℓ_2	$\top$

Note that μ_h is not $\otimes$-transitive, since $\mu_h(b, x_2) \otimes \mu_h(x_2, x_1)$ and $\mu_h(b, x_1)$ are not comparable.

Concerning $\otimes$-$\approx_B$-antisymmetry, if a fuzzy relation ρ_B is $\otimes$-$\approx_B$-antisymmetric then any other relation μ such that $\mu \leq \rho_B$ is also $\otimes$-$\approx_B$-antisymmetric. If there were a fuzzy adjunction $(i, h)\colon \langle X, \approx_B, \rho_X\rangle \leftrightharpoons \langle B, \approx_B, \rho_B\rangle$ for a contraction h and a suitable fuzzy relation ρ_B, we would have $\mu_h^t \leq \rho_B$, and then μ_h^t is antisymmetric. Hence, the $\otimes$-$\approx_B$-antisymmetry of μ_h^t is a necessary condition.

Lemma 2 *Let $X \neq \varnothing$ be a subset of B such that $\langle X, \approx_B, \rho_X\rangle$ is a fuzzy preordered structure and let $h\colon B \to X$ be a contraction. Then, the h-reflexive closure of ρ_X, that is μ_h, satisfies the following properties:*

1. *$\mu_h(b_1, b_2) \leq \mu_h^2(b_1, b_2)$ for any $b_1, b_2 \in B$.*
2. *$\mu_h^2(x, b) = \mu_h(x, b)$ for any $x \in X$ and $b \in B$.*
3. *$\mu_h^2(b_1, b_2) = \mu_h^3(b_1, b_2)$ for any $b_1, b_2 \in B$.*
4. *μ_h^2 is the transitive closure of μ_h.*

Theorem 2 *Let X be a nonempty subset of a fuzzy structure $\mathcal{B} = \langle B, \approx_B\rangle$ together with any $\approx_B$-reflexive, $\approx_B$-$\otimes$-antisymmetric and $\otimes$-transitive fuzzy relation ρ_X on $\mathcal{X} = \langle X, \approx_B \rangle$. Consider a contraction $h\colon B \to X$ and the h-reflexive closure μ_h, then the following statements are equivalent:*

1. *There exists a $\approx_B$-reflexive, $\approx_B$-$\otimes$-antisymmetric and $\otimes$-transitive fuzzy relation ρ_B in $\mathcal{B}$ such that the pair (i, h) is a fuzzy adjunction between $\langle \mathcal{X}, \rho_X\rangle$ and $\langle \mathcal{B}, \rho_B\rangle$.*

 ISBN: 978-84-608-6082-2

2. μ_h^2 *is* $\otimes$-$\approx_B$*-antisymmetric.*

According to Theorems 2 and 3 (below), the necessary and sufficient condition for the existence of a fuzzy preorder structure on B and the right adjoint for the natural inclusion $i\colon X \to B$, for a subset X of B is the existence of a contraction $h\colon B \to X$ such that μ_h^2 is antisymmetric.

Finally, we identify suitable conditions that guarantee this antisymmetry.

Theorem 3 *Let X be a nonempty subset of a fuzzy structure $\mathcal{B} = \langle B, \approx_B \rangle$ together with any $\approx_B$-reflexive, $\approx_B$-$\otimes$-antisymmetric and $\otimes$-transitive fuzzy relation ρ_X on X. Consider a contraction $h\colon B \to X$ and the h-reflexive closure μ_h, then μ_h^2 is $\otimes$-$\approx_B$-antisymmetric if and only if the following conditions hold:*

(i) $\rho_X(x, h(b)) \leq \bigwedge_{y \in X} (b \approx_B y) \otimes \rho_X(y, x) \to (b \approx_B x)$ *for all* $x \in X$ *and* $b \notin X$.

(ii) $(b_1 \approx_B x) \otimes \rho_X(x, h(b_2)) \otimes (b_2 \approx_B y) \otimes \rho_X(y, h(b_1)) \leq (b_1 \approx_B b_2)$*, for all* $x, y \in X$ *and* $b_1, b_2 \in B \setminus X$.

Acknowledgements

Partially supported by Spanish Ministry of Science projects TIN2014-59471-P and TIN2015-70266-C2-1-P, co-funded by the European Regional Development Fund (ERDF).

References

[1] R. Bělohlávek. Fuzzy Galois connections. *Mathematical Logic Quarterly*, 45(4):497–504, 1999.

[2] R. Bělohlávek. Lattices of fixed points of fuzzy Galois connections. *Mathematical Logic Quarterly*, 47(1):111–116, 2001.

[3] U. Bodenhofer, B. De Baets and J. Fodor. A compendium of fuzzy weak orders: Representations and constructions. *Fuzzy Sets and Systems* 158(8):811–829, 2007.

[4] F. Börner. Basics of Galois connections. *Lect. Notes in Computer Science*, 5250:38–67, 2008.

[5] I.P. Cabrera, P. Cordero, F. García-Pardo, M. Ojeda-Aciego, and B. De Baets. On the construction of adjunctions between a fuzzy preposet and an unstructured set. *Submitted*, 2016.

[6] I.P. Cabrera, P. Cordero, B. De Baets, F. García-Pardo, and M. Ojeda-Aciego. On the existence of right adjoints for surjective mappings between fuzzy structures. In Proc. of *Concept Lattices and their Applications, CLA*, 2016. To appear.

[7] K. Denecke, M. Erné, and S. L. Wismath. *Galois connections and applications*, Kluwer Academic Publishers, 2004.

[8] Y. Djouadi and H. Prade. Interval-valued fuzzy Galois connections: Algebraic requirements and concept lattice construction. *Fundamenta Informaticae*, 99(2):169–186, 2010.

[9] J. G. García, I. Mardones-Pérez, M. A. de Prada-Vicente, and D. Zhang. Fuzzy Galois connections categorically. *Mathematical Logic Quarterly*, 56(2):131–147, 2010.

[10] F. García-Pardo, I.P. Cabrera, P. Cordero, and M. Ojeda-Aciego. On the construction of fuzzy Galois connections. *Proc. of XVII Spanish Conference on Fuzzy Logic and Technology*, pages 99-102, 2014.

[11] F. García-Pardo, I.P. Cabrera, P. Cordero, M. Ojeda-Aciego, and F.J. Rodríguez. On the definition of suitable orderings to generate adjunctions over an unstructured codomain. *Information Sciences* 286:173-187, 2014.

[12] G. Georgescu and A. Popescu. Non-commutative fuzzy Galois connections. *Soft Computing*, 7(7):458–467, 2003.

[13] J. Järvinen. Pawlak's information systems in terms of Galois connections and functional dependencies. *Fundamenta Informaticae*, 75:315–330, 2007.

[14] S. Kuznetsov. Galois connections in data analysis: Contributions from the soviet era and modern russian research. *Lect. Notes in Computer Science*, 3626:196–225, 2005.

[15] A. Melton, D. A. Schmidt, and G. E. Strecker. Galois connections and computer science applications. *Lect. Notes in Computer Science*, 240:299–312, 1986.

[16] S.-C. Mu and J. Oliveira. Programming from Galois connections. *Journal of Logic and Algebraic Programming*, 81(6):680–704, 2012.

[17] J. Propp. A Galois connection in the social network. *Mathematics Magazine*, 85(1):34–36, 2012.

[18] M. Wolski. Galois connections and data analysis. *Fundamenta Informaticae*, 60:401–415, 2004.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Coating of C_{60} by para-H_2 and ortho-D_2: revisiting the solvation shell

Florent Calvo[1] **and Ersin Yurtsever**[2]

[1] *Univ. Grenoble Alpes, LIPHY, F-38000 Grenoble, CNRS, LiPhy, F-38000 Grenoble*

[2] *Koç University, Rumelifeneriyolu, Sariyer, Istanbul 34450, Turkey,*

emails: florent.calvo@univ-grenoble-alpes.fr, eyurtsev@edu.ku.tr

Abstract

The solvation of Buckminsterfullerene C_{60} by para-hydrogen and ortho-deuterium clusters has been modeled using a dedicated potential and path-integral molecular dynamics simulations at low temperature (2 K). The solvation shell obtained from the distribution of radial distances is found to be complete near 50 molecules, in agreement with recent mass spectrometry measurements. Deuteration increases the shell size by one, indicating a denser shell owing to less prominent vibrational delocalization for this heavier isotope.

Key words: Hydrogen clusters; fullerenes; quantum solvation; path-integral molecular dynamics

1 Introduction

The adsorption of hydrogen on carbonaceous molecules or extended organic substrates such as graphite or graphene has attracted considerable attention, especially fuelled in the recent years by the possibility of exploiting these light compounds for reversible storage of hydrogen [1, 2, 3]. The corrugation of graphitic surfaces contributes to the formation of ordered layers of hydrogen, the first layer being $\sqrt{3} \times \sqrt{3}$ commensurate with the honeycomb lattice of graphite [4]. In finite clusters of carbon such as fullerenes, this corrugation is also present and even strengthened by the curvature of these roughly spherical compounds. The adsorption of hydrogen on fullerenes was successfully reported in mass spectrometry experiments by the Scheier group, using the cryogenic medium of helium droplets [5]. These authors reported the formation of two ion series, namely $(H_2)_n C_{60}^+$ and $(H_2)_n HC_{60}^+$ and found a strong drop in

 ISBN: 978-84-608-6082-2

mass abundances at $n = 32$. However, shell completion was reported to occur significantly later at $n = 49$. This number differs significantly from early path-integral Monte Carlo simulations carried by Turnbull and Boninsegni [6] who obtained shell sizes of 32 or 40 for neutral C_{60} depending on the commensurate nature of the coating layer. Another puzzling issue with the measurements carried out by Kaiser and coworkers [5] is the absence of deuteration effects on the reported shell size, whether deuteration is known to affect the phase diagram of hydrogen absorbed on graphite quite significantly.

In the present contribution, we have revisited the problem of hydrogen coating on fullerenes, focusing here on the simplest case of C_{60} and leaving for future investigation the specific effects of charge and fullerene size or shape. Our computational study relies on quantum simulations of hydrogen clusters around the fullerene, treating interactions explicitly and accounting for nuclear quantum effects using the framework of path-integral molecular dynamics (PIMD). Our results indicate a shell size of 50 for para-hydrogen and 51 for ortho-deuterium, in relatively good agreement with experimental data and suggesting indeed minor deuteration effects.

2 Potential energy surface

We consider a C_{60} Buckminsterfullerene molecule rigidly fixed at the center of mass of the reference frame and a set of n hydrogen molecules adsorbed exohedrally around it. PIMD simulations were carried out to determine the quantum mechanical equilibrium distribution of the hydrogen molecules at the cryogenic temperature of $T = 2$ K, which is above the quantum melting transition known to take place in pure para-H_2 clusters [7]. The para-H_2 and ortho-D_2 molecules lie in their rovibrational $J = 0$ state, and at the low densities and low pressures experienced in gas-phase compounds can be assumed to rotate freely and behave like effective pointlike particles. This approximation allows us to employ the Silvera-Goldman (SG) pairwise potential [8] to describe the interactions between such molecules, following here most earlier theoretical investigations on related systems [6, 7, 9]. The spherical approximation for hydrogen molecules implies a similar description of interactions between H_2 and the carbon atoms of the fullerene. Here we use a SG form for this potential as well and an additive expression for the total energy $V_{\rm tot}$ of the n-molecule system:

$$\begin{aligned} V_{\rm tot} &= \sum_{i\in H_2} V_i, \\ V_i &= \sum_{j\in C_{60}} \left[V_{\rm rep}(r_{ij}) + V_{\rm disp}(r_{ij})\right], \end{aligned}$$

where the repulsive and additive contributions have the same form as the original SG potential. In the case of an hydrocarbon dopant molecule, separate contributions from hydrogen

 ISBN: 978-84-608-6082-2

and carbon atoms have to be considered, as well as a polarization contribution to account for the presence of partial charges on the different atoms with different electronegativities. Here for the neutral and chemically homogeneous C_{60}, we neglect these charges and the polarization term altogether.

The potential was parametrized against electronic structure calculations for an H_2 molecule approaching small aromatic hydrocarbons such as benzene [10]. These calculations were performed using the spin-component-scaled (SCS) MP2 method [11] with the MOLPRO software package [12]. The final potential energy curves were averaged over different orientations of the H_2 molecule. We show in Fig. 1 the resulting potential energy curve for H_2 approaching C_{60}, compared to other existing potentials in the literature of the Lennard-Jones (LJ) type [13, 14], or to a few electronic structure data at the levels of density-functional theory (wB97xD and PBE0 functionals) [5] or symmetry-adapted perturbation theory (SAPT) [15]. In the latter case, only the values of the binding energies at the reported equilibrium distances are indicated. Clearly our potential performs very satis-

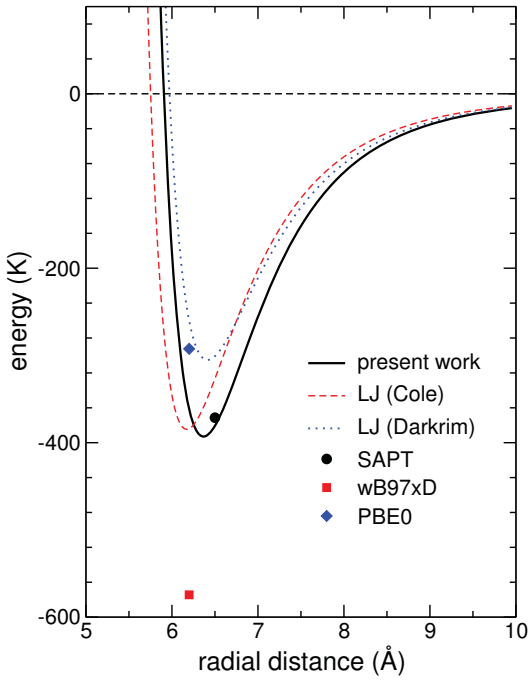


Figure 1: Potential energy between the H_2 and C_{60} molecule as a function of the distance between their centers of mass. In addition to the present potential (solid lines), the graphs shows the predictions of three electronic structure calculations (symbols [5, 15]) and of two Lennard-Jones potentials as dashed (Cole [13]) and dotted (Darkrim [14]) lines.

 ISBN: 978-84-608-6082-2

factorily against those alternative models, and even manages to reconcile the LJ potential of Stan and Cole [13] with the highly accurate SAPT calculation of Korona *et al.* [15]. As noted by the original authors [5], the PBE0 (wB97xD) calculations are likely to understimate (overestimate) the true binding energies, and this is reflected in their quite significant deviation, although the equilibrium distance near 6.2 Å remains close to our prediction.

3 Path-integral molecular dynamics simulations

Based on the potential energy surface described in the previous section, we have first determined the energetically most stable structures of $(H_2)_nC_{60}$ clusters, fixing once for all the number of hydrogens $n = 100$ to be large enough to exceed the completion of the first solvation shell. The basin-hopping method was used to locate the putative global minima of the cluster, taking a maximum number of 10^4 local minimizations and a temperature of $T = 10$ K in the Metropolis acceptance step. The PIMD trajectories were initiated from the most stable structure thus obtained, and integrated over 100 ps using a time step of 0.2 fs and a Trotter number of $P = 256$ at $T = 2$ K. From the atomic positions of each bead describing the polymer necklace, the radial distributions of hydrogen molecules around the center of C_{60} were integrated, and the simulations were repeated for the two isotopes of para-hydrogen and ortho-deuterium only differing in the particle masses.

These radial distributions are shown in Fig. 2 for the two systems of $(\text{para-}H_2)_{100}C_{60}$ and $(\text{ortho-}D_2)_{100}C_{60}$. The distributions both exhibit a clear two-peak structure indicative

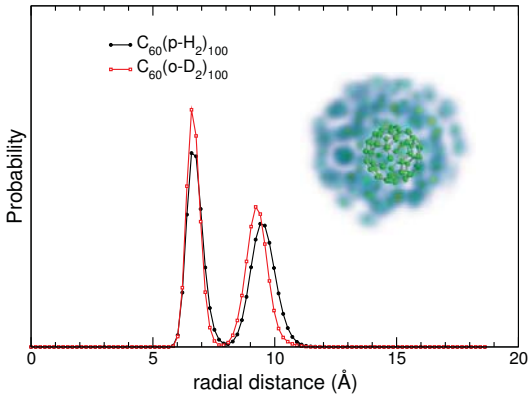


Figure 2: Radial distributions of H_2 molecules around C_{60}, as obtained from path-integral molecular dynamics simulations of $(H_2)_{100}C_{60}$ at 2 K, distinguishing para-H_2 and ortho-D_2 isotopes. The three-dimensional integrated density is also depicted as an inset, using the Mayavi2 suite of python scripts [17].

of a complete first shell surrounded by an at least partially filled second shell. Delocaliza-

 ISBN: 978-84-608-6082-2

tion effects have several manifestations on these graphs. First, the two peaks are rather broad, but the first shell appears noticeably thinner than the second shell. This behavior is consistent with the stronger interaction of hydrogen to C_{60} than to itself (by a factor of approximately 13), which tends to pin the hydrogen atoms closer to the fullerene surface. The thickness of the first shell, as measured by the width at mid height of the first peak in the radial distribution, is also nearly unchanged by deuteration, which further suggests nearly that hydrogen behaves nearly classically in contact with the fullerene.

Although of similar thickness, the first shells of the two isotopic systems do not accomodate exactly the same number of molecules. Integrating the radial distributions over the first peak region yields shell sizes of 50 and 51 for para-H_2 and ortho-D_2, respectively. These values are very close to the experimental result [5] of 49 for the C_{60}^+ dopant, at least much closer than the value of 40 obtained from the quantum Monte Carlo calculations of Turnbull and Boninsegni [6]. This discrepancy between the two computational results likely originates from different methodologies and notably the slightly different potentials employed by the latter authors and their use of the grand-canonical ensemble.

Another manifestation of delocalization effects on the radial distributions is the generally higher surface density of ortho-deuterium layers, consistently with the correspondingly smaller de Broglie thermal wavelength for the heavier isotope. The second layer for ortho-D_2 is not only more compact, but also slightly closer to the fullerene center. A three-dimensional density plot obtained from integrating the PIMD trajectory, shown as an inset in Fig. 2, sheds more light onto the two-shell structure of the hydrogen solvent, and notably reveals that the second shell is incomplete.

A residual difference of one (para-H_2) or two (ortho-D_2) molecules exists between our prediction of the size of the first solvation shell and the experimental measurements. Several explanations can be proposed for this difference, starting with the possible influence of the charge since mass spectra are recorded on ionic species. Owing to polarization forces, binding of H_2 to C_{60}^+ could be significantly enhanced relative to the neutral system. Such stronger forces could bring the first layer closer to the fullerene surface and one molecule could be ejected from the new optimal structure. It could also be that due to such stronger forces the neglect of rotational structure for H_2 molecules be no longer valid especially in contact with the fullerene. More sophisticated all-atom potentials would then be necessary in order to circumvent this difficulty. Finally, temperature effects could also be invoked to explain the difference of one or two molecules between the measurements and our calculation. The mass spectra of Ref. [5] for ionic complexes solvated by hydrogen were recorded in after further embedding into helium droplets, thereby imposing a temperature lower than 2 K. It is possible that at such very low temperatures the hydrogen solvation shell adopts a different structure due to entropic effects that are further competing with zero-point effects [16]. It would be interesting to address such an issue by extending the present calculations at 0.5 K, for instance, much closer to the experimental situation of true helium droplets.

 ISBN: 978-84-608-6082-2

4 Concluding remarks

Despite exhibitting a high symmetry, C_{60} buckminsterfullerene coated by hydrogen shows a complex solvation structure close to 50 molecules in the first shell that cannot be explained based on simple geometric arguments known in the case of flat graphitic surfaces. Our quantum simulations reproduce fairly well the experimentally measured value but, contrary to experiment, suggest some noticeable effects of deuteration.

Beyond the present purely structural investigation, a wealth of information can be analysed from the PIMD trajectories and would be worth processing further. In particular, the extent of vibrational delocalization, the possibility of quantum melting at lower temperature, its possible quenching in presence of the dopant and the energetics of solvation would all be valuable to persue in more details. Comparison with the case of helium, where intricate finite size effects have been reported below the completion of the first solvation shell [18], would also be of interest in the broader framework of solvation by quantum gases.

Acknowledgements

Financial support through a joint grant from CNRS/TUBITAK is gratefully acknowledged. Work done in Istanbul has been supported by the COST/CM1410 action project TUBITAK/Z214Z07.

References

[1] H. Furukawa and O. M. Yaghi, *Storage of hydrogen, methane, and carbon dioxide in highly porous covalent organic frameworks for clean energy applications*, J. Am. Chem. Soc. **131** (2009) 8875–8883.

[2] P. Jena, *Materials for hydrogen storage: past, present, and future*, J. Phys. Chem. Lett. **2** (2011) 206–211.

[3] N. Park, K. Choi, J. Hwang, D. W. Kim, D. O. Kim and J. Ihm, *Progress on first-principles-based materials design for hydrogen storage*, Proc. Natl. Acad. Sci. U.S.A. **109** (2012) 19893–19899.

[4] L. W. Bruch, M. W. Cole and E. Zaremba, *Physical Adsorption*, Mineola, NY: Dover, 2007.

[5] A. Kaiser, C. Leidlmair, P. Bartl, S. Zöttl, S. Denifl, A. Mauracher, M. Probst, P. Scheier, and O. Echt, *Adsorption of hydrogen on neutral and charged fullerene: Experiment and theory*, J. Chem. Phys. **138** (2013) 074311.

 ISBN: 978-84-608-6082-2

[6] J. D. Turnbull and M. Boninsegni, *Adsorption of* para*-hydrogen on fullerenes*, Phys. Rev. B **71** (2005) 205421.

[7] F. Mezzacapo and M. Boninsegni, *On the possible "supersolid" character of parahydrogen clusters*, J. Phys. Chem. A **115** (2011) 6831–6837.

[8] I. F. Silvera and V. V. Goldman, *The isotropic intermolecular potential for H_2 and D_2 in the solid and gas phases*, J. Chem. Phys. **69** (1978) 4209–4213.

[9] M. C. Gordillo, *H_2 on corrugated graphene: Diffusion Monte Carlo calculations*, Phys. Rev. B **88** (2013) 041406(R).

[10] F. Calvo and E. Yurtsever, *Solvation of carbonaceous molecules by para-H_2 and ortho-D_2 clusters. Polycyclic aromatic hydrocarbons*, J. Chem. Phys. (2016) in press.

[11] S. Grimme, *Improved second-order Møller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies*, J. Chem. Phys. **118** (2003) 9095–9102.

[12] MOLPRO (version 2012.1) is a package of ab initio programs written by H.-J. Werner, J.-J. Knowles, F. R. Manby, M. Schütz, P. Celani, G. Knizia, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut *et al.*

[13] G. Stan and M. W. Cole, *Hydrogen adsorption in nanotubes*, J. Low Temp. Phys. **110** (1998) 539–544.

[14] D. Levesque, A. Gicquel, F. L. Darkrim and S. B. Kayiran, *Monte Carlo simulations of hydrogen storage in carbon nanotubes*, J. Phys. Cond. Matt. **14** (2002) 9285–9293.

[15] T. Korona, A. Hesselmann, and H. Dodziuk, *Symmetry-adapted perturbation theory applied to endohedral fullerene complexes: A stability study of $H_2@C_{60}$ and $2H_2@C_{60}$*, J. Chem. Theory Comput. **5** (2009) 1585–1596.

[16] F. Calvo, J. P. K. Doye and D. J. Wales, *Quantum partition functions from classical distributions: Application to rare-gas clusters*, J. Chem. Phys. **114** (2001) 7312–7329.

[17] P. Ramachandran and G. Varoquaux, *Mayavi: a package for 3D visualization of scientific data*, IEEE Computing in Science & Engineering **13** (2011) 40–51.

[18] F. Calvo, *Size-induced melting and reentrant freezing in fullerene-doped helium clusters*, Phys. Rev. B **85** (2012) 060502(R).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Exact solutions and conservation laws of a Generalized Fornberg-Whitham Equation

J.C. Camacho[1] **and M.S. Bruzón**[1]

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

emails: josecarlos.camacho@uca.es, m.bruzon@uca.es

Abstract

The application of Lie transformations group theory for the construction of solutions of nonlinear partial differential equations is one of the most active fields of research in the theory of nonlinear partial differential equations and applications.

In this paper we consider the generalized Fornberg-Whitham Equation (GFWE)

$$\Delta \equiv u_t - u_{xxt} + \beta u_x = u u_{xxx} - \alpha u^n u_x + 3 u_x u_{xx}, \tag{1}$$

and we study the values of constants for which equation (1) admits the classical symmetry group.

The symmetry group of the equation (1) will be generated by a vector field of the form

$$X = \xi(x,t,u)\partial_x + \tau(x,t,u)\partial_t + \phi(x,t,u)\partial_u.$$

We require that the infinitesimal generator leaves invariant the set of solutions of the equation, and only obtain, in the general case, the group of space and time translations,

$$\mathbf{v}_1 = \partial_x, \qquad \mathbf{v}_2 = \partial_t,$$

For some special choices of the constants it can be extended in the cases listed below
Case 1. $\alpha = 0$.

$$\mathbf{v}_{31} = \beta t \partial_x + t\partial_t - (u - \beta)\partial_u.$$

Case 2. $n = 1$.

$$\mathbf{v}_{32} = t\partial_x - \frac{\alpha - 1}{\beta} t\partial_t + (\frac{\alpha - 1}{\beta} u - 1)\partial_u.$$

In order to reduce the equation (1) to ODEs with one independent variable, we construct the similarity variables and similarity forms of field variables. By using the characteristic equations

$$\frac{dx}{\xi(x,t,u)} = \frac{dt}{\tau(x,t,u)} = \frac{du}{\phi(x,t,u)}$$

 ISBN: 978-84-608-6082-2

we obtain the similarity variables and the reduced ODEs.
In the general case, the infinitesimal generator of the optimal system is $\mathbf{v}_1 + \lambda\mathbf{v}_2$. Next, the similarity solutions are

$$z = x - \lambda t,$$

$$u(x,t) = h(z).$$

Finally, equation (1) is reduced into ODE

$$\lambda\,(h''' - h') - (h\,h')'' - \frac{\alpha}{n+1}\,(h^{n+1})' + \beta\,h' = 0. \tag{2}$$

We obtain travelling wave solutions of equation (2), and hence, we get exact travelling wave solutions of equation (1).
From generators$\lambda\mathbf{v}_1 + \mathbf{v}_{31}$ and $\lambda\mathbf{v}_1 + \mathbf{v}_{32}$ we obtain the reduced equations.

In order to obtain conservation laws for equation (1) we apply the multiplier method [1, 2, 3, 4]. Conservation laws play a important role in the resolution of problems in which certain physical properties do not change in the course of time. In [5] was proven that existence of a large number of conservation laws of a partial differential equations is indicative of the integrability of the equation. The conservation laws are also derived for equation (1).

In [3] the authors gave a general treatment of a direct conservation law method for partial differential equations expressed in a standard Cauchy-Kovaleskaya form. By applying this procedure we obtain the conservation laws

$$\begin{aligned} C^1 &= u_{xx} - u. \\ C^2 &= -\frac{1}{2}u^2 + (u_{xx} - 1)u + \frac{1}{2}u_x^2. \end{aligned} \tag{3}$$

In the other hand, in [6] the authors investigated the nonlinear self-adjointness of equation (1) by using the Ibragimovs conservation theorem [7], and they claim that only the trivial conservation laws are extractable. So, by the multiplier method we obtain new conservation laws.

Acknowledgements

The authors acknowledge the financial support from Junta de Andalucía group FQM–201 and Universidad de Cádiz.

References

[1] S. C. Anco, G. Bluman, *Direct constrution of conservation laws from field equations*, Phys. Rev. Lett. **78** (1997) 2869–2873.

[2] S. C. Anco, G. Bluman, *Direct constrution method for conservation laws of partial differential equations Part I:Examples of conservation law classifications*, Eur. J. Appl. Math. **13** (2002) 545–566.

[3] S. C. Anco, G. Bluman, *Direct constrution method for conservation laws for partial differential equations Part II: General treatment*, Euro. Jnl of Applied Mathematics **41** (2002) 567–585.

[4] S. C. Anco, , *Symmetries and conservation laws of the generalized Krichever-Novikov equation* arXiv: 1407.1258v4 [nlin.SI], 2015.

[5] G. W. Bluman, S. Kumei, , *Symmetries and Differential Equations* Applied Mathematical Sciences, 81, Springer-Verlag, New York, 1989.

[6] M. S. Hashemi, A. Haji-Badali, P. Vafadar, *Group Invariant Solutions and Conservation Laws of the Fornberg Whitham Equation*, Z. Naturforsch. **69a** (2014) 489–496. DOI: 10.5560/ZNA.2014-0037

[7] N. H. Ibragimov, *Transformation Groups Applied to Mathematical Physics*, Reidel–Dordrecht, 1985.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

The implicit midpoint method for the modified anomalous sub-diffusion equation with a nonlinear source term

Xuenian Cao[1] and Xianxian Cao[1]

[1] *School of Mathematics and Computation,, Xiangtan University, Xiangtan,Hunan 411105, China*

emails: cxn@xtu.edu.cn,

Abstract

In this paper, the implicit midpoint method is used for the numerical solution of the semi-discrete modified anomalous sub-diffusion equation with a nonlinear source term, and the weighted and shifted $Grünwald - Letnikov$ difference operator and the compact difference operator are applied to approximate the $Riemann - Liouville$ fractional derivative and space partial derivative respectively, then the numerical scheme is constructed. The stability and the convergence of this method are analyzed. Numerical experiment is given which check the accuracy of this method and confirm our theoretical results.

Key words: Modified anomalous sub-diffusion equation; implicit midpoint method; compact difference operator; stability; convergence.

1 Introduction

There has been increasing interest in the description of physical and chemical processes by means of equations involving fractional derivatives over the last decades [1, 2, 3, 4, 5, 6, 7, 8]. Recently there are models that have been proposed to describe process that become less anomalous as time progresses by the inclusion of a secondary fractional time derivative acting on a diffusion operator with a nonlinear source term[15, 16]

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} = (A_0 D_t^{1-\alpha} + B_0 D_t^{1-\beta}) \frac{\partial^2 u(x,t)}{\partial x^2} + f(u(x,t),x,t) \\ 0 \le x \le S, 0 \le t \le T \\ u(0,t) = \varphi_1(t), u(S,t) = \varphi_2(t), 0 \le t \le T \\ u(x,0) = \psi(x), 0 \le x \le S \end{aligned} \tag{1}$$

 ISBN: 978-84-608-6082-2

where $0 < \alpha, \beta < 1$, A, B are positive constants,the symbols ${}_0D_t^{1-\alpha}u(x,t)$,${}_0\,D_t^{1-\beta}u(x,t)$ are the *Riemann − Liouville* fractional derivative operator, which are defined by

$$ {}_0D_t^{1-\gamma}u(x,t) = \frac{1}{\Gamma(\gamma)}\frac{\partial}{\partial t}\int_0^t \frac{u(x,\eta)}{(t-\eta)^{1-\gamma}}d\eta, \quad \gamma = \alpha, \beta, $$

where $\Gamma(\cdot)$ is the gamma function, $f(u,x,t)$ satisfies the *Lipschitz* condition with respect u:

$$ |f(u,x,t) - f(\upsilon,x,t)| \leq L|u-\upsilon|, \forall u, \upsilon, $$

where L is a *Lipschitz* condition.

Assume the problem (1) has a unique sufficiently smooth solution $u(x,t)$.

Much work have been done on developing numerical methods for solving the modified anomalous sub-diffusion equation , see e.g.([9, 11, 12, 14, 15, 16]). However, effective numerical methods and their numerical analysis are still too little.

The outline of this paper is as follows. In section 2, the numerical method for the modified anomalous sub-diffusion equation is given. Then, in section 3, stability and convergence analysis are investigated, respectively. Section 4 is used to present numerical results, comparing the fixed stepsize implementation on a test problem. Numerical experiment shows that the proposed method is high accuracy and efficiency for solving the modified anomalous sub-diffusion equation .

2 Numerical method

In this paper, we assume that $u(x,t) \in U(\Omega)$, where

$$ U(\Omega) = \{u(x,t) \mid \frac{\partial^6 u(x,t)}{\partial x^6}, \frac{\partial^3 u(x,t)}{\partial x^2 \partial t}, \frac{\partial^3 u(x,t)}{\partial t^3} \in C(\Omega)\}, $$

whereas $\Omega = \{(x,t)|0 \leq x \leq S,\ 0 \leq t \leq T\}$.

For the space interval $[0,S]$ and time interval $[0,T]$, we choose the grid points as follows $x_j = jh,\ j = 0,1,\cdots,M,\quad t_n = n\tau,\ n = 0,1,\cdots,N$, where $h = \frac{S}{M}$ denotes spatial step size, $\tau = \frac{T}{N}$ denotes time stepsize. The exact solution and numerical solution at the point (x_j,t_n) are denoted by $u(x_j,t_n)$ and u_j^n respectively.

Lemma 2.1. *If $u(x,t) \in U(\Omega)$, then*

$$ (1 + \frac{1}{12}\delta_x^2)\frac{\partial^2 u(x_j,t_n)}{\partial x^2} = \frac{\delta_x^2 u(x_j,t_n)}{h^2} + O(h^4) \tag{2} $$

where $\delta_x^2 u(x_j,t_n) = u(x_{j-1},t_n) - 2u(x_j,t_n) + u(x_{j+1},t_n)$.

Lemma 2.2. *[13, 14]Let $f(t) \in L^1(\mathbb{R}), {}_{-\infty}D_t^{\alpha+2}$ and its Fourier transform belong to $L^1(\mathbb{R})$,and define the weighted and shifted Grünwald − Letenikov difference operator by*

$$D^{\alpha}_{\tau,p,q}f(t) = \frac{\alpha-2q}{2(p-q)}A^{\alpha}_{\tau,p}f(t) + \frac{2p-\alpha}{2(p-q)}A^{\alpha}_{\tau,q}f(t).$$

then we have

$$D^{\alpha}_{\tau,p,q}f(t) =_{-\infty} D^{\alpha}_t f(t) + O(\tau^2), t \in \mathbb{R}$$

where p, q are integers and $p \neq q$,$A^{\alpha}_{\tau,r}$ is the Grünwald − Letenikov approximation to the Riemann-Liouville fractional derivative by

$$A^{\alpha}_{\tau,r}f(t) = \tau^{-\alpha}\sum_{k=0}^{\infty} g_k^{(\alpha)} f(t-(k-r)\tau), r = p, q,$$

where $g_0^{(\alpha)} = 1, g_k^{(\alpha)} = (1-\frac{\alpha+1}{k})g_{k-1}^{(\alpha)}, k \geq 1$.

Applying the implicit midpoint formula to solve the equation (1) at $x = x_j, t = t_n$, we have

$$\begin{aligned} u(x_j,t_{n+1}) = u(x_j,t_n) &+ \tau(A_0D_t^{1-\alpha} + B_0D_t^{1-\beta})\frac{\partial^2}{\partial x^2}(\frac{u(x_j,t_{n+1})+u(x_j,t_n)}{2}) \\ &+ \tau f(\frac{u(x_j,t_{n+1})+u(x_j,t_n)}{2}), x_j, t_{n+\frac{1}{2}}) + O(\tau^3), \end{aligned} \tag{3}$$

$$j = 1, 2, \cdots, M-1, n = 0, 1, \cdots, N-1.$$

In Lemma 2.2, we choose $(p, q) = (0, -1)$,then

$$\begin{aligned} {}_0D_t^{1-\gamma}(\frac{\partial^2 u(x_j,t_n)}{\partial x^2}) &= \tau^{\gamma-1}\sum_{i=0}^{n}\lambda_i^{(1-\gamma)}\frac{\partial^2 u(x_j,t_{n-i})}{\partial x^2} + O(\tau^2), \\ {}_0D_t^{1-\gamma}(\frac{\partial^2 u(x_j,t_{n+1})}{\partial x^2}) &= \tau^{\gamma-1}\sum_{i=0}^{n+1}\lambda_i^{(1-\gamma)}\frac{\partial^2 u(x_j,t_{n+1-i})}{\partial x^2} + O(\tau^2), \end{aligned} \tag{4}$$

where

$$\lambda_0^{(1-\gamma)} = \frac{3-\gamma}{2}g_0^{(1-\gamma)}, \lambda_i^{(1-\gamma)} = \frac{3-\gamma}{2}g_i^{(1-\gamma)} + \frac{\gamma-1}{2}g_{i-1}^{(1-\gamma)}, \quad i \geq 1, \gamma = \alpha, \beta.$$

Insert (4) in (3), we get that

$$\begin{aligned} u(x_j,t_{n+1}) = u(x_j,t_n) &+ \sum_{i=0}^{n+1}(\frac{A\tau^{\alpha}}{2}\mu_i^{(1-\alpha)} + \frac{B\tau^{\beta}}{2}\mu_i^{(1-\beta)})\frac{\partial^2 u(x_j,t_{n+1-i})}{\partial x^2} \\ &+ \tau f(\frac{u(x_j,t_{n+1})+u(x_j,t_n)}{2}), x_j, t_{n+\frac{1}{2}}) + O(\tau^3). \end{aligned} \tag{5}$$

where

$$\mu_0^{(1-\gamma)} = \lambda_0^{(1-\gamma)}, \mu_i^{(1-\gamma)} = \lambda_i^{(1-\gamma)} + \lambda_{i-1}^{(1-\gamma)}, i \geq 1, \gamma = \alpha, \beta.$$

Furthermore, applying operator $(1 + \frac{1}{12}\delta_x^2)$ premultiplication both sides of Eq.(5), and Lemma 2.1, we obtain

$$\begin{aligned}(1 + \frac{1}{12}\delta_x^2)u(x_j, t_{n+1}) = (1 + \frac{1}{12}\delta_x^2)u(x_j, t_n) + \sum_{i=0}^{n+1} \varpi_i \delta_x^2 u(x_j, t_{n+1-i}) \\ + \tau(1 + \frac{1}{12}\delta_x^2) f(\frac{u(x_j, t_{n+1}) + u(x_j, t_n)}{2}), x_j, t_{n+\frac{1}{2}}) + R_j^{n+1}\end{aligned} \tag{6}$$

where

$$\varpi_i = \frac{A\tau^\alpha}{2h^2}\mu_i^{(1-\alpha)} + \frac{B\tau^\beta}{2h^2}\mu_i^{(1-\beta)} \tag{7}$$

$$R_j^{n+1} = O(h^4)\sum_{i=0}^{n+1}(\frac{A\tau^\alpha}{2}\mu_i^{(1-\alpha)} + \frac{B\tau^\beta}{2}\mu_i^{(1-\beta)}) + (1 + \frac{1}{12}\delta_x^2)O(\tau^3). \tag{8}$$

So the numerical scheme for solving the problem (1) is devised as follows

$$\begin{aligned}(1 + \frac{1}{12}\delta_x^2)u_j^{n+1} = (1 + \frac{1}{12}\delta_x^2)u_j^n + \sum_{i=0}^{n+1} \varpi_i \delta_x^2 u_j^{n+1-i} + \tau(1 + \frac{1}{12}\delta_x^2) f_j^{n+\frac{1}{2}}. \\ n = 0, 1, \cdots, N-1, j = 1, 2, \cdots, M-1. \\ u_j^0 = 0, j = 0, 1, \cdots, M, \\ u_0^n = \varphi_1(t_n), u_M^n = \varphi_2(t_n), n = 0, 1, \cdots, N.\end{aligned} \tag{9}$$

where $f_j^{n+\frac{1}{2}} = f(\frac{u_j^n + u_j^{n+1}}{2}, x_j, t_{n+\frac{1}{2}})$.

3 Stability and convergence analysis

Let

$$\Upsilon = \{\xi | \xi = (\xi_0, \xi_1, \cdots, \xi_M)\}, \widehat{\Upsilon} = \{\xi | \xi \in \Upsilon, \xi_0 = \xi_M = 0\}.$$

For any $u, \upsilon \in \Upsilon$, we denote

$$\Delta u_i = u_{i+1} - u_i, \Delta \upsilon_i = \upsilon_{i+1} - \upsilon_i.$$

and further we define for any $u, \upsilon \in \widehat{\Upsilon}$

$$\begin{aligned}&(u, \upsilon) = h\sum_{i=1}^{M-1} u_i \upsilon_i, \| u \|^2 = (u, u), \\ &\langle \Delta u, \Delta \upsilon \rangle = h\sum_{i=0}^{M-1} \Delta u_i \Delta \upsilon_i, \| \Delta u \|_d^2 = \langle \Delta u, \Delta \upsilon \rangle, \\ &\delta_x^2 u = (0, \delta_x^2 u_1, \delta_x^2 u_2, ..., \delta_x^2 u_{M-1}, 0).\end{aligned}$$

Lemma 3.1. *let $\omega_1, \omega_2 \geq 0, \widetilde{\tau} > 0, \eta_0, \eta_1, \cdots, \eta_{\widetilde{N}}$ are a series of nonnegative real numbers,which satisfying*

$$\eta_m \leq \omega_2 + \omega_1 \widetilde{\tau} \sum_{i=0}^{m-1} \eta_i, m = 0, 1, \cdots, \widetilde{N},$$

then

$$\eta_m \leq \omega_2 e^{\omega_1 m \widetilde{\tau}}, m = 0, 1, \cdots, \widetilde{N}.$$

Lemma 3.2. *[14] $\{\varpi_i\}_{i=0}^{\infty}$ are defined as in (7),then for any positive integer k and real vector $(\upsilon_1, \upsilon_2, \cdots, \upsilon_k)^T \in R^k$,then*

$$\sum_{n=0}^{k-1} (\sum_{i=0}^{n} \varpi_i \upsilon_{n+1-i}) \upsilon_{n+1} \geq 0.$$

Lemma 3.3. *[14] For $\upsilon \in \widehat{\Upsilon}$, we have*

$$< \delta_x^2 u, \upsilon >= - < \Delta u, \Delta \upsilon > .$$

Lemma 3.4. *[12] For $\upsilon \in \widehat{\Upsilon}$, we have*

$$\parallel \Delta \upsilon \parallel_d^2 \leq 4 \parallel \upsilon \parallel^2 .$$

Lemma 3.5. *For $u, \upsilon \in \Upsilon$, we have*

$$(i) \parallel u \parallel^2 - (u, \upsilon) = \frac{1}{2}(\parallel u \parallel^2 - \parallel \upsilon \parallel^2) + \frac{h}{2} \sum_{i=1}^{M-1} (u_i - \upsilon_i)^2.$$

$$(ii) \parallel \Delta u \parallel_d^2 - \langle \Delta u, \Delta \upsilon \rangle = \frac{1}{2}(\parallel \Delta u \parallel_d^2 - \parallel \Delta \upsilon \parallel_d^2) + \frac{h}{2} \sum_{i=1}^{M-1} (\Delta u_i - \Delta \upsilon_i)^2.$$

Theorem 3.1. *Assuming $0 < \tau \leq \tau_0 = \frac{4-\upsilon}{9L}$, where $\upsilon \in (0, 1)$ is any given positive number, the implicit midpoint method defined by (9) is stable, i.e. there exists a positive constant $C = C(\alpha, \beta, \upsilon, \varpi_0, L, T)$ such that*

$$\max_{0 \leq n \leq N} \parallel E^n \parallel \leq C \parallel E^0 \parallel, 0 < \tau \leq \tau_0,$$

where $E^n = (0, e_1^n, e_2^n, \cdots, e_{M-1}^n, 0)^T, e_j^n = u_j^n - \widetilde{u_j^n}$, $\widetilde{u_j^n}$ denotes the parallel approximation solution for the numerical method (9) starting from another initial values.

Theorem 3.2. *Assuming $0 < \tau \leq \tau_0 = \frac{4-\upsilon}{3(2+3L)}$, where $\upsilon \in (0, 1)$ is any given positive number,the implicit midpoint method defined by (9) is convergent such that*

$$\max_{0 \leq n \leq N} \parallel \eta^n \parallel = O(h^4 + \tau^2),$$

where $\eta^n = [0, \eta_1^n, \eta_2^n, \cdots, \eta_{M-1}^n, 0]^T, \eta_j^n = u(x_j, t_n) - u_j^n$.

4 Numerical experiment

Let

$$E_\infty(h,\tau) = \max_{0\le k\le N}\max_{0\le i\le M} \mid u(x_i,t_k) - u_i^k \mid,$$

where $u(x_i,t_k)$ and u_i^k denote the exact solution and the numerical solution, respectively.

Observed order of the numerical method about spatial variable and temporal variable are denoted as follow

$$Order_\tau = log_2(\frac{E_\infty(h,2\tau)}{E_\infty(h,\tau)}),$$

$$Order_h = log_2(\frac{E_\infty(2h,4\tau)}{E_\infty(h,\tau)}),$$

where h is spatial stepsize, τ is time stepsize.

Example consider the following nonlinear problem

$$\frac{\partial u(x,t)}{\partial t} = \frac{1}{2}(\frac{\partial^{1-\alpha}}{\partial t^{1-\alpha}} + \frac{\partial^{1-\beta}}{\partial t^{1-\beta}})\frac{\partial^2 u(x,t)}{\partial x^2} + f(u(x,t),x,t) \tag{10}$$

with the boundary and initial conditions

$$\begin{aligned} u(0,t) = t^2, u(S,t) = t^2 e^S, 0 < t \le T, \\ u(x,0) = 0, 0 \le x \le S, \end{aligned} \tag{11}$$

where the nonlinear source term

$$f(u(x,t),x,t) = u^3(x,t) + e^x(2t - t^6 e^{2x} - \frac{t^{1+\alpha}}{\Gamma(2+\alpha)} - \frac{t^{1+\beta}}{\Gamma(2+\beta)}). \tag{12}$$

The exact solution of the problem (10)-(12) is

$$u(x,t) = t^2 e^x.$$

For comparison purpose, we use our method (9) (Im) and the method presented in [9] (Ec) to solve the problem (10)-(12), the numerical results are listed in Table 4.1-4.2.

Table4.1 Errors and the time observed orders of numerical methods for $S = 1, T = 1, \alpha = 0.35, \beta = 0.65, h = \frac{1}{32}$

τ	$E_\infty(h,\tau)(Im)$	$Order_\tau(Im)$	$E_\infty(h,\tau)(Ec)$	$Order_\tau(Ec)$
$\frac{1}{20}$	*6.1476e-4*	-	*2.8970e-3*	-
$\frac{1}{40}$	*1.5349e-4*	*2.0019*	*1.4893e-3*	*0.9599*
$\frac{1}{80}$	*3.8380e-5*	*1.9997*	*7.5492e-4*	*0.9802*
$\frac{1}{160}$	*9.5994e-6*	*1.9993*	*3.8004e-4*	*0.9901*
$\frac{1}{320}$	*2.4000e-6*	*1.9999*	*1.9066e-4*	*0.9952*
$\frac{1}{640}$	*5.9948e-7*	*2.0013*	*9.5496e-5*	*0.9975*

Table4.2 Errors and the space observed order of numerical method(Im) for $S = 1, T = 1, \alpha = 0.95, \beta = 0.15$

h, τ	$E_\infty(h, \tau)$	$Order_h$
$h = \frac{1}{4}, \tau = \frac{1}{128}$	*9.5262e-6*	-
$h = \frac{1}{8}, \tau = \frac{1}{512}$	*6.2991e-7*	*3.9187*
$h = \frac{1}{16}, \tau = \frac{1}{2048}$	*3.7420e-8*	*4.0733*

From Table 4.1 and Table 4.2, we see that the observed convergence order of our method (9) is nearly $O(h^4 + \tau^2)$, which means the convergence order of the numerical results match that the theoretical result. In addition, from Table 4.1 we can see that our method has higher accuracy than the method presented in [9].

Acknowledgements

This work is supported by projects from the National Natural Science Foundation of China(No.11271311 and No.11171282).

References

[1] A A Kilbas, H M Srivastava, J J Trujillo . Theory and applications of fractional differential equations. Elsevier Science Limited, 2006.

[2] I. Petras Fractional-order nonlinear systems: modeling, analysis and simulation. Springer, 2011.

[3] I. Podlubny. Fractional differential equations. Academic Press, 1998.

[4] Koeller, R.C. Application of fractional calculus to the theory of viscoelasticity. J. Appl. Mech. 51.229C307 (1984).

[5] Becker-Kern, P., Meerschaert, M.M., Scheffler, H.P. Limit theorem for continuous-time random walks with two time scales. J. Appl. Prob. 41, 455C466 (2004).

[6] Meerschaert, M.M., Zhang, Y., Baeumerc, B. Particle tracking for fractional diffusion with two time scales. Comput. Math. Appl. 59, 1078C1086 (2010).

[7] Gorenflo, R., Mainardi, F., Scalas, E., Raberto, M. Fractional calculus and continuous-time finance.III, The diffusion limit. In: Mathematical Finance, Trends in Math, pp. 171C180. Birkh?user, Basel(2001).

[8] Meerschaert, M.M., Scalas, E. Coupled continuous time random walks in finance. Physica A 370,114C118 (2006).

[9] A. Mohebbi, M. Abbaszadeh, M. Dehghan. A high-order and unconditionally stable scheme for the modified anomalous fractional sub-diffusion equation with a nonlinear source term. J. Comput.Phys. 240 (2013) 36-48.

[10] F. Liu, C. Yang, K. Burrage. Numerical method and analytical technique of the modified anomalous subdiffusion equation with a nonlinear source term. J. Comput. Appl. Math. 231 (2009) 160-176.

[11] Y.Chen,Chang-Ming. Numerical scheme with high order accuracy for solving a modified fractional diffusion equation. Applied Mathematics and Computation 224(2014)772-782.

[12] Y.F.Li,D.L.Wang. Improved efficient difference method for the modified anomalous subdiffusion equation with a nonlinear source term. International Journal of Computer Mathematics,2016.

[13] W. Y. Tian, H. Zhou, and W. H. Deng. A class of second order difference approximations for solving space fractional diffusion equations. Math. Comp. doi: 10.1090/S0025-5718- 2015-02917-2 (2015).

[14] Z.B Wang, Seakweng Vong. Compact difference schemes for the modified anomalous fractional sub-diffusion equation and the fractional diffusion-wave equation. J.Comput.Phys. 277 (2014) 1-15.

[15] F. Liu, C. Yang, K. Burrage. Numerical method and analytical technique of the modified anomalous subdiffusion equation with a nonlinear source term. J. Comput. Appl. Math. 231 (2009) 160C176.

[16] Q. Liu, F. Liu, I. Turner, V. Anh. Finite element approximation for a modified anomalous subdiffusion equation. Appl. Math. Model. 35 (2011) 4103C4116.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The correlation attack to LFSRs as a syndrome decoding problem

Sara D. Cardell[1], Joan-Josep Climent[2] and Alicia Roca[3]

[1] *Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas*

[2] *Departament de Matemàtiques, Universitat d'Alacant*

[3] *Departamento de Matemática Aplicada, Universidad Politécnica de Valencia*

emails: sdcardell@ime.unicamp.br, jcliment@ua.es, aroca@mat.upv.es

Abstract

One of the most successful attacks against a secret random sequence of bits produced by certain minimum-length linear feedback shift registers (LFSRs) has been achieved by the fast correlation attack by Meier and Staffelbach (1988, 1989). Correlation attacks are often viewed as decoding problems. Assume that a sequence $\boldsymbol{y}$ produced by an LFSR is sent through a transmission channel. Let $\boldsymbol{z}$ be the received channel output, which is correlated to the sequence $\boldsymbol{y}$ with correlation probability $1-\varepsilon$ with $0.25 \leq \varepsilon \leq 0.5$.

A natural way of analyzing the stream of bits produced by a minimum-length LFSR is to understand it as an autonomous system. Therefore, the sequence $\boldsymbol{y}$ can be interpreted as a codeword in the binary $[n,k]$-code $\mathcal{C}$ generated by the corresponding observability matrix. Then, the problem of the attacker can be reformulated as follows: Given a received word $\boldsymbol{z}$, find the transmitted codeword $\boldsymbol{y}$.

Taking advantage of these approach, we analyze the fast correlation attack as a syndrome decoding problem. We propose a decoding algorithm based on the representation technique of the syndromes by Becker, Joux, May and Meurer (2012).

Key words: LFSR, correlation attack, keystream sequence, companion matrix, autonomous system, syndrome decoding problem, decoding representation technique.
MSC 2000: 68P25, 68P30, 94A60, 94B35

1 Introducttion

The sequences generated by linear feedback shift registers (LFSRs) have properties desirable for keystreams. However, with the Berlekamp-Massey algoritm [7] it is very easy to compute the feedback polynomial of the LFSR given at least $2L(\boldsymbol{y})$ bits of the sequence $\boldsymbol{y}$, where $L(\boldsymbol{y})$ is the linear complexity of the sequence $\boldsymbol{y}$. We need to destroy the linearity before the sequence can be used as keystream. One classical approach is to combine several binary LFSRs via a nonlinear Boolean function. In practical stream cipher systems (like the Geffe generator) it is often found that a correlation occurs between the keystream and the output of an individual LFSR within the key generator. Among the different kinds of attacks against stream ciphers, correlation attacks are one of the most important. Correlation attacks are a class of plaintext attacks for breaking stream ciphers whose keystream is generated by combining the output of several LFSRs. They were first introduced by Siegenthaler [13] and are based on a model where the keystream is viewed as a noisy version of the output of some of the constituent LFSRs; it is assumed that the noise is additive and independent of the underlying LFSR sequence. Correlation attacks exploit a statistical weakness that arises from a poor choice of the Boolean function. Meier and Staffelbach [9] presented two different algorithms for fast correlation attacks, using a correlation between the keystream and the output stream of an LFSR. In the past years several algorithms for correlation attacks and fast correlation attacks where proposed (see, to mention only a few examples, [1, 3, 4, 5, 6, 8, 11, 12]).

2 Statement of the problem

Let $\mathbb{F}_2$ be the Galois field of two elements. Assume that

$$f(\mathsf{x}) = c_0 + c_1\mathsf{x} + c_2\mathsf{x}^2 + \cdots + c_{k-1}\mathsf{x}^{k-1} + \mathsf{x}^k \in \mathbb{F}_2[\mathsf{x}]$$

is the feedback polynomial of an LFSR and consider the **companion matrix**

$$A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & c_0 \\ 1 & 0 & 0 & \cdots & 0 & c_1 \\ 0 & 1 & 0 & \cdots & 0 & c_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & c_{k-2} \\ 0 & 0 & 0 & \cdots & 1 & c_{k-1} \end{bmatrix} \in \mathbb{F}_2^{k\times k}$$

of $f(\mathsf{x})$ and the column matrix

$$C = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}^T \in \mathbb{F}_2^{k\times 1}.$$

 ISBN: 978-84-608-6082-2

A way to describe the LFSR is by means of the autonomous system

$$\left.\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t A \\ y_t &= \boldsymbol{x}_t C \end{aligned}\right\} \qquad t = 0, 1, 2 \ldots \tag{1}$$

where $\boldsymbol{x}_0$ is the **initial state** of the system.

If $\boldsymbol{x}_0 = (y_0, y_1, \ldots, y_{k-1})$ is the initial state, then the t-th stream bit y_t, for $t \geq k$, can be computed using expression (1). Moreover, if $f(\mathsf{x})$ is a primitive polynomial, then the output sequence $y_0, y_1, \ldots, y_{k-1}, y_k, \ldots$ has maximal period $2^k - 1$.

Let n be a positive integer such that $k < n < 2^k - 1$. We can compute the output sequence $\boldsymbol{y} = (y_0, y_1, y_2, \ldots, y_{k-1}, y_k, y_{k+1}, \ldots, y_{n-1})$ as $\boldsymbol{y} = \boldsymbol{x}_0 G$ where

$$G = \begin{bmatrix} C & AC & A^2C & A^3C & \cdots & A^{n-2}C & A^{n-1}C \end{bmatrix}$$

is the **observability matrix** of the system given by expression (1).

Assume that we do not know neither the sequence $\boldsymbol{y}$ nor the initial state $\boldsymbol{x}_0$. Assume also that we know the sequence

$$\boldsymbol{z} = (z_0, z_1, z_2, \ldots, z_{k-1}, z_k, z_{k+1}, \ldots, z_{n-1})$$

which is correlated to the sequence $\boldsymbol{y}$ with correlation probability $1 - \varepsilon$ (usually $0.25 \leq \varepsilon \leq 0.5$). The idea of the correlation attack is to view the sequence $\boldsymbol{z}$ as a perturbation of the sequence $\boldsymbol{y}$ by a binary symmetric memoryless noise channel with $Pr(z_i = y_i) = 1 - \varepsilon$ (see [9, 10]). Thus the LFSR sequence $\boldsymbol{y}$ is interpreted as a codeword in the $[n, k]$-code $\mathcal{C}$ generated by matrix G and the keystream sequence $\boldsymbol{z}$ as the received channel output. The correlation attack can now be reformulated as: *Given a received word $\boldsymbol{z}$, find the transmitted codeword $\boldsymbol{y}$*. This means that $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{e}$, where $\boldsymbol{e} \in \mathbb{F}_2^n$ is a vector with Hamming weight $\mathrm{wt}(\boldsymbol{e}) = \varepsilon n$. So, the error capability ω of the code $\mathcal{C}$ is given by $w = \varepsilon n$.

Assume now that $n = mk$ for some positive integer m. Since $k < n < 2^k - 1$, it follows that $1 < m < \frac{2^k - 1}{k}$. Then

$$\begin{bmatrix} C & AC & A^2 & A^3 & \cdots & A^{k-1}C \end{bmatrix} = I_k$$

where I_k denotes the $k \times k$ identity matrix. Consequently

$$\begin{bmatrix} A^kC & A^{k+1}C & A^{k+2}C & A^{k+3}C & \cdots & A^{2k-1}C \end{bmatrix} \\ = A^k \begin{bmatrix} C & AC & A^2 & A^3 & \cdots & A^{k-1}C \end{bmatrix} = A^k.$$

So, we can write the generator matrix G of $\mathcal{C}$ as

$$G = \begin{bmatrix} I_k & A^k & A^{2k} & A^{3k} & \cdots & A^{(m-2)k} & A^{(m-1)k} \end{bmatrix}$$

 ISBN: 978-84-608-6082-2

which is in systematic form. Consequently, if $B = (A^k)^T$, then the corresponding parity-check matrix H of $\mathcal{C}$ is given by

$$H = \begin{bmatrix} B & I_k & & & & \\ B^2 & & I_k & & & \\ B^3 & & & I_k & & \\ \vdots & & & & \ddots & \\ B^{m-2} & & & & & I_k \\ B^{m-1} & & & & & \end{bmatrix}. \tag{2}$$

3 The decoding algorithm

Now, assume that

$$\boldsymbol{y} = (\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{m-2}, \boldsymbol{y}_{m-1}) \quad \text{and} \quad \boldsymbol{z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{m-2}, \boldsymbol{z}_{m-1}) \tag{3}$$

where

$$\left.\begin{array}{ll} \boldsymbol{y}_i & = (y_{ik}, y_{ik+1}, y_{ik+2}, \ldots, y_{(i+1)k-2}, y_{(i+1)k-1}) \\ \boldsymbol{z}_i & = (z_{ik}, z_{ik+1}, z_{ik+2}, \ldots, z_{(i+1)k-2}, z_{(i+1)k-1}) \end{array}\right\} \text{for } i = 0, 1, 2, \ldots, m-2, m-1.$$

Since $\boldsymbol{y} \in \mathcal{C}$, we have that

$$H\boldsymbol{y}^T = \boldsymbol{0}^T$$

and therefore

$$B^i \boldsymbol{y}_0^T + \boldsymbol{y}_i^T = \boldsymbol{0}^T, \quad \text{for} \quad i = 1, 2, \ldots, m-1. \tag{4}$$

This means that $(\boldsymbol{y}_0, \boldsymbol{y}_i)$ is a codeword of the $[2k, k]$-code $\mathcal{C}_i$ whose parity-check matrix is $H_i = \begin{bmatrix} B^i & I_k \end{bmatrix}$, for $i = 1, 2, \ldots, m-1$.

On the other hand, let $\boldsymbol{s} \in \mathbb{F}_2^{(m-2)k}$ be the syndrome of $\boldsymbol{z}$, i.e., $H\boldsymbol{z}^T = \boldsymbol{s}^T$, or equivalently,

$$B^{i+1} \boldsymbol{z}_0^T + \boldsymbol{z}_{i+1}^T = \boldsymbol{s}_i^T, \quad \text{for} \quad i = 0, 1, 2, \ldots, m-2$$

where

$$\boldsymbol{s}_i = (s_{ik}, s_{ik+1}, s_{ik+2}, \ldots, s_{(i+1)k-2}, s_{(i+1)k-1}), \quad \text{for} \quad i = 0, 1, 2, \ldots, m-2.$$

Since $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{e}$, we have that

$$\boldsymbol{z}_i = \boldsymbol{y}_i + \boldsymbol{e}_i, \quad \text{for } i = 0, 1, 2, \ldots, m-1, \tag{5}$$

and consequently,

$$B^{i+1} \boldsymbol{e}_0^T + \boldsymbol{e}_{i+1}^T = \boldsymbol{s}_i^T, \quad \text{for} \quad i = 0, 1, 2, \ldots, m-2. \tag{6}$$

Note that if $\boldsymbol{s}_{i_0} = \mathbf{0}$ for some $i_0 \in \{0, 1, 2, \ldots, m-2\}$, then $(\boldsymbol{z}_0, \boldsymbol{z}_{i_0+1}) = (\boldsymbol{y}_0, \boldsymbol{y}_{i_0+1})$, and we can compute $\boldsymbol{y}$ from expressions (3), (4) and (5). Therefore, from now on, we will assume that $\boldsymbol{s}_i \neq \mathbf{0}$, for $i = 0, 1, 2, \ldots, m-2$.

Now, expression (6) suggests the following brute force attack.

Algorithm 1:

Input: *matrix H, vector $\boldsymbol{s}$, and integer w*

Output: *vector $\boldsymbol{e} \in \mathbb{F}_2^n$ such that $H\boldsymbol{e}^T = \boldsymbol{s}^T$ and $\mathrm{wt}(\boldsymbol{e}) = w$, and the codeword $\boldsymbol{y}$*

1. *For each $\boldsymbol{e}_0 \in \mathbb{F}_2^k$, compute*

$$\boldsymbol{e}_{i+1}^T = \boldsymbol{s}_i^T + B^{i+1}\boldsymbol{e}_0^T, \quad \textit{for} \quad i = 0, 1, 2, \ldots, m-2$$

   and consider the vector $\boldsymbol{e} = (\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_{m-1})$.

2. *If $\mathrm{wt}(\boldsymbol{e}) = w$, assume that $\boldsymbol{e}$ is the error vector and then compute the corresponding codeword as $\boldsymbol{y} = \boldsymbol{z} + \boldsymbol{e}$.*

Since the previous attack is infeasible for $k \geq 32$, we will use a modification of the method proposed by Becker, Joux, May, and Meurer [2].

To improve the calculations, we introduce additional hypotheses on the weight distribution of the vector $\boldsymbol{e} \in \mathbb{F}_2^n$. Hence, we search for vectors $\tilde{\boldsymbol{e}}$ that can be decomposed into $\tilde{\boldsymbol{e}} = (\tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1, \tilde{\boldsymbol{e}}_2) \in \mathbb{F}_2^k \times \mathbb{F}_2^k \times \mathbb{F}_2^{n-2k}$ where $\mathrm{wt}(\tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1) = p$ and $\mathrm{wt}(\tilde{\boldsymbol{e}}_2) = w - p$, and $H\tilde{\boldsymbol{e}}^T = \boldsymbol{s}^T$. Then, as a consequence of expression (2), after obtaining the vector $(\tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1)$ such that

$$\boldsymbol{s}_0^T = Q_1 \begin{bmatrix} \tilde{\boldsymbol{e}}_0^T \\ \tilde{\boldsymbol{e}}_1^T \end{bmatrix}, \quad \text{where} \quad Q_1 = \begin{bmatrix} B & I_k \end{bmatrix},$$

we can obtain the vector $\tilde{\boldsymbol{e}}_2$ as $\tilde{\boldsymbol{e}}_2^T = Q_2\tilde{\boldsymbol{e}}_0^T + \tilde{\boldsymbol{s}}_1^T$ where

$$\tilde{\boldsymbol{s}}_1 = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{m-2}) \qquad \text{and} \qquad Q_2 = \begin{bmatrix} B^2 \\ B^3 \\ \vdots \\ B^{m-2} \\ B^{m-1} \end{bmatrix}.$$

Therefore we focus our attention on efficiently computing the vector $(\tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1)$. Following [2] we propose the algorithm described below, which is based on the next ideas.

In turn, to find $(\tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1) \in \mathbb{F}_2^k \times \mathbb{F}_2^k$ such that $\mathrm{wt}((\tilde{\boldsymbol{e}}_0, \tilde{\boldsymbol{e}}_1)) = p$ we proceed in different steps.

First of all, we search for two collections of vectors $\tilde{\boldsymbol{u}} \in \mathbb{F}_2^{2k}$ of half the target weight, such that the nonzero components of every vector of one of the collections appear in disjoint

positions with respect to the nonzero components of the vectors of the second one. This is obtained in the step 3 of Algorithm 2.

Second, we find "matches" between vectors of the two previously obtained collection of vectors, i.e. when multiply Q_1 by the vectors of the first and on those of the second collections, a prescribed number of its components "overlap", in the sense that for the prescribed coordinates the difference of the results is a constant randomly selected vector. Taking advantage of it, we obtain a new collection of vectors with certain knowledge about their last components. The process is repeated in steps 6 and 7 of Algorithm 2, leading to a final collection of vectors forced to produce the vector $\boldsymbol{s}_0$, i.e., the first part of the syndrome vector $\boldsymbol{s}$, when multiply Q_1 by each of such vectors.

Algorithm 2:

Input: *Q_1, $\boldsymbol{s}_0$ and p with $0 < p \leq 2k$.*

Output: *A set $\mathcal{L}$ of vectors $\boldsymbol{u} \in \mathbb{F}_2^{2k}$ such that $\mathrm{wt}(\boldsymbol{u}) = p$ and $Q_1\boldsymbol{u}^T = \boldsymbol{s}_0^T$.*

1. *Define $p_1 = \frac{p}{2} + \varepsilon_1$ and $p_2 = \frac{p_1}{2} + \varepsilon_2$ for some ε_1 and ε_2 such that p_1 and p_2 are even numbers.*

2. *For $i = 1, 2, 3, 4$, choose random partitions $\mathcal{P}_i^{(1)}, \mathcal{P}_i^{(2)}$ of $\{1, 2, \ldots, 2k\}$ to create the basic sets*

$$\mathcal{B}_i^{(1)} = \left\{ \boldsymbol{u} \in \mathbb{F}_2^{2k} \mid \mathrm{wt}(\boldsymbol{u}) = \frac{p_2}{2} \text{ and } u_l = 0 \text{ for all } l \in \mathcal{P}_i^{(2)} \right\},$$
$$\mathcal{B}_i^{(2)} = \left\{ \boldsymbol{u} \in \mathbb{F}_2^{2k} \mid \mathrm{wt}(\boldsymbol{u}) = \frac{p_2}{2} \text{ and } u_l = 0 \text{ for all } l \in \mathcal{P}_i^{(1)} \right\}.$$

3. *Choose a random integer r_1 and a random vector $\boldsymbol{t}_1^{(1)} \in \mathbb{F}_2^{r_1}$, and set $\boldsymbol{t}_2^{(1)} = \boldsymbol{s}_0^* + \boldsymbol{t}_1^{(1)}$, where $\boldsymbol{s}_0^*$ denotes the last r_1 components of $\boldsymbol{s}_0$.*

4. *Choose a random integer r_2 and two random vectors $\boldsymbol{t}_1^{(2)}, \boldsymbol{t}_3^{(2)} \in \mathbb{F}_2^{r_2}$, and set $\boldsymbol{t}_2^{(2)} = \boldsymbol{t}_1^{(1)*} + \boldsymbol{t}_1^{(2)}$ and $\boldsymbol{t}_4^{(2)} = \boldsymbol{t}_2^{(1)*} + \boldsymbol{t}_3^{(2)}$, where $\boldsymbol{t}_1^{(1)*}$ and $\boldsymbol{t}_2^{(1)*}$ denote the last r_2 components of $\boldsymbol{t}_1^{(1)}$ and $\boldsymbol{t}_2^{(1)}$ respectively.*

5. *For $i = 1, 2, 3, 4$, use the basic sets $(B_i^{(1)}, B_i^{(2)})$ and the vector $\boldsymbol{t}_i^{(2)}$ to define the set*

$$\mathcal{L}_i^{(2)} = \left\{ \boldsymbol{u}^{(2)} \in \mathbb{F}_2^{2k} \mid \mathrm{wt}\left(\boldsymbol{u}^{(2)}\right) = p_2 \text{ and } \left(Q_1(\boldsymbol{u}^{(2)})^T\right)^* = (\boldsymbol{t}_i^{(2)})^T \right\}$$

where $\left(Q_1(\boldsymbol{u}^{(2)})^T\right)^*$ denotes the last r_2 components of $(Q_1\boldsymbol{u}^{(2)})^T$.

6. *For $j = 1, 2$, use the sets $(L_{2j-1}^{(2)}, L_{2j}^{(2)})$ and the vector $\boldsymbol{t}_j^{(2)}$ to define the set*

$$\mathcal{L}_j^{(1)} = \left\{ \boldsymbol{u}^{(1)} \in \mathbb{F}_2^{2k} \mid \mathrm{wt}\left(\boldsymbol{u}^{(1)}\right) = p_1 \text{ and } \left((Q_1\boldsymbol{u}^{(1)})^T\right)^* = (\boldsymbol{t}_j^{(1)})^T \right\}$$

where $\left(Q_1\boldsymbol{u}^{(1)}\right)^*$ denotes the last r_1 components of $Q_1\boldsymbol{u}^{(1)}$.

7. *Use the sets* $(L_1^{(1)}, L_2^{(1)})$ *and the vector* $\boldsymbol{s}_0$ *to define the set*

$$\mathcal{L} = \left\{ \boldsymbol{u} \in \mathbb{F}_2^{2k} \mid \mathrm{wt}(\boldsymbol{u}) = p \text{ and } Q_1 \boldsymbol{u}^T = \boldsymbol{s}_0^T \right\}.$$

Acknowledgements

The first author was supported by FAPESP with number of process 2015/07246-0. The second and third authors are partially supported by grant MIMECO MTM2015-68805-REDT. The third author is also partially supported by grant MINECO MTM2013-40960-P.

References

[1] M. Ågren, C. Löndahl, M. Hell, and T. Johansson. A survey on fast correlation attacks. *Cryptography and Communications*, 4(3–4):173–202, 2012.

[2] A. Becker, A. Joux, A. May, and A. Meurer. Decoding random binary linear codes in $2^{n/20}$: How $1 + 1 = 0$ improves information set decoding. In D. Pointcheval and T. Johansson, editors, *Advances in Cryptology – EUROCRYPT 2012*, volume 7237 of *Lecture Notes in Computer Science*, pages 520–536. Springer-Verlag, Berlin, 2012.

[3] A. Canteaut and M. Naya-Plasencia. Correlation attacks on combination generators. *Cryptography and Communications*, 4(3–4):147–171, 2012.

[4] V. V. Chepyzhov, T. Johansson, and B. Smeets. A simple algorithm for fast correlation attacks on stream ciphers. In B. Schneier, editor, *Fast Software Encryption – FSE 2000*, volume 1978 of *Lecture Notes in Computer Science*, pages 181–195. Springer-Verlag, Berlin, 2001.

[5] T. Johansson and F. Jönsson. Theoretical analysis of a correlation attack based on convolutional codes. *IEEE Transactions on Information Theory*, 48(8):2173–2181, 2002.

[6] P. Lu and L. Huang. A new correlation attack on LFSR squences with high error tolerance. *Progress in Computer Science and Applied Logic*, 23:67–83, 2004.

[7] J. L. Massey. Shift-register synthesis and BCH decoding. *IEEE Transactions on Information Theory*, 15(1):122–127, 1969.

[8] W. Meier. Fast correlation attacks: methods and countermeasures. In A. Joux, editor, *Fast Software Encryption – FSE 2011*, volume 6733 of *Lecture Notes in Computer Science*, pages 55–67. Springer-Verlag, Berlin, 2011.

[9] W. Meier and O. Staffelbach. Fast correlation attacks on stream ciphers. In C. G. Günter, editor, *Advances in Cryptology – EUROCRYPT'88*, volume 330 of *Lecture Notes in Computer Science*, pages 301–314. Springer-Verlag, Berlin, 1988.

[10] W. Meier and O. Staffelbach. Fast correlation attacks on certain stream ciphers. *Journal of Cryptology*, 1(3):159–176, 1989.

[11] H. Molland, J. E. Mathiassen, and T. Helleseth. Improved fast correlation attack using low rate codes. In K. G. Paterson, editor, *Cryptography and Coding*, volume 2898 of *Lecture Notes in Computer Science*, pages 67–81. Springer-Verlag, New York, NY, 2003.

[12] M. Noorkami and F. Fekri. A fast correlation attack via unequal error correcting LDPC codes. In T. Okamoto, editor, *Topics in Cryptology – CT-RSA 2004*, volume 2964 of *Lecture Notes in Computer Science*, pages 54–46. Springer-Verlag, Berlin, 2004.

[13] T. Siegenthaler. Decrypting a class of stream ciphers using ciphertext only. *IEEE Transactions on Computers*, 34(1):81–85, 1985.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On a simple construction of primitive polynomials

Sara D. Cardell[1] **and Joan-Josep Climent**[2]

[1] *Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Brazil*

[2] *Departament de Matemàtiques, Universitat d'Alacant, Spain*

emails: sdcardell@ime.unicamp.br, jcliment@ua.es

Abstract

In this work we present a construction of primitive polynomials over $\mathbb{F}_p$ based on the isomorphism between $\mathbb{F}_{p^b}$ and $\mathbb{F}_p[C]$, where C is the companion matrix of a primitive polynomial of degree b in $\mathbb{F}_p$.

Key words: companion matrix, primitive polynomial, ring isomorphism

Primitive polynomials have been extensively studied because of their important applications (see, for example, [6]). They are widely used in cryptographic applications such that pseudo-random sequence generation. For example, every linear feedback shift register (LFSR) with maximum period is built from a primitive polynomial [3].

Various tables of primitive polynomials over finite fields were presented in the technical literature [1, 4]. Primitive polynomials over the binary field, $\mathbb{F}_2$, have received particular attention, due to their use in the generation of linear recurring sequences widely employed in testing, coding theory, cryptography, communication systems, and many other areas of electrical engineering [8, 9, 7].

Let $\mathbb{F}_p$ be the Galois field of p elements, with p a positive prime integer. A generator of the cyclic group $\mathbb{F}_p^*$ is called a **primitive element** of $\mathbb{F}_p$.

A polynomial $A(\mathsf{x}) \in \mathbb{F}_p[x]$ of degree $m \geq 1$ is called **primitive** over $\mathbb{F}_p$ if it is the minimal polynomial over $\mathbb{F}_p$ of a primitive element of $\mathbb{F}_{p^m}$. Thus, a primitive polynomial over $\mathbb{F}_p$ of degree m may be described as a monic polynomial that is irreducible over $\mathbb{F}_p$ and has a root $\alpha \in \mathbb{F}_{p^m}$ that generates the multiplicative group of $\mathbb{F}_{p^m}$.

The **companion matrix** of a monic polynomial $A(\mathsf{x}) = a_0 + a_1\mathsf{x} + a_2\mathsf{x}^2 + \cdots + a_{m-1}\mathsf{x}^{m-1} + \mathsf{x}^m \in \mathbb{F}_p[\mathsf{x}]$ is given by the $m \times m$ matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -a_{m-2} \\ 0 & 0 & \cdots & 1 & -a_{m-1} \end{bmatrix}.$$

Some authors define the companion matrix of $A(x)$ as $\mathbf{A}^T$.

We can see the elements in $\mathbb{F}_{p^m}$ as matrices. Consider the primitive polynomial $U(\mathsf{x}) = u_0 + u_1\mathsf{x} + u_2\mathsf{x}^2 + \cdots + u_{m-1}\mathsf{x}^{m-1} + \mathsf{x}^m \in \mathbb{F}_p[\mathsf{x}]$. For this purpose, we consider the companion matrix $\mathbf{U}$ of the primitive polynomial $U(\mathsf{x})$. In this case, it is well-known that $\mathbb{F}_p[\mathbf{U}] = \{F(\mathbf{U}) \mid F(\mathsf{x}) \in \mathbb{F}_p[\mathsf{x}]\}$ is a field which is isomorphic to $\mathbb{F}_{p^m}$ (see, for example [5]). Now, we can consider the field isomorphism $\psi : \mathbb{F}_{p^m} \longrightarrow \mathbb{F}_p[\mathbf{U}]$ given by $\psi(\alpha) = \mathbf{U}$, where $\alpha \in \mathbb{F}_{p^m}$ is a primitive element, and $\psi(0) = \mathbf{O}$ (see [2, 5]). Then we can write

$$\mathbb{F}_{p^m} = \left\{\mathbf{O}, \mathbf{I}, \mathbf{U}, \mathbf{U}^2, \ldots, \mathbf{U}^{p^m-2}\right\}.$$

This isomorphism can be extended to the following ring isomorphism (see [2]):

$$\begin{array}{rcl} \Psi : \mathrm{Mat}_{r\times r}(\mathbb{F}_{p^m}) & \longrightarrow & \mathrm{Mat}_{r\times r}(\mathbb{F}_p[\mathbf{U}]) \\ \mathbf{A} = [a_{i,j}] & \mapsto & \Psi(\mathbf{A}) = [\psi(a_{i,j})] \end{array} \tag{1}$$

Next theorem shows how to use this ring isomorphism in order to construct primitive polynomials with coefficients in $\mathbb{F}_p$ using primitive polynomials with coefficients in $\mathbb{F}_{p^m}$.

Theorem 1 *Let $\mathbf{V}$ be the companion matrix of a primitive polynomial $V(\mathsf{x}) \in \mathbb{F}_{p^m}[\mathsf{x}]$ of degree r, and let Ψ be the ring isomorphism given in expression (1). Then, the characteristic polynomial $C(\mathsf{x}) = \det(\mathsf{x}\mathbf{I} - \Psi(\mathbf{V})) \in \mathbb{F}_p[\mathsf{x}]$ is a primitive polynomial of degree rm.*

Example 1 *Consider the primitive polynomial $U(\mathsf{x}) = 1 + \mathsf{x} + \mathsf{x}^2 \in \mathbb{F}_2[\mathsf{x}]$. We can use this polynomial to construct the Galois field $\mathbb{F}_{2^2}$ of 4 elements. For this purpose, we consider the companion matrix $\mathbf{U}$ of $U(\mathsf{x})$, that is,*

$$\mathbf{U} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

and, since $U(\mathsf{x})$ is a primitive polynomial, we know that

$$\mathbb{F}_{2^2} \approx \mathbb{F}_2[\mathbf{U}] = \{\mathbf{O}, \mathbf{I}, \mathbf{U}, \mathbf{U} + \mathbf{I}\} = \{\mathbf{O}, \mathbf{I}, \mathbf{U}, \mathbf{U}^2\}.$$

Now, as we saw before, given a primitive element $\alpha \in \mathbb{F}_{2^2}$ *with* $\alpha^2+\alpha+1=0$*, we can consider the map* $\psi : \mathbb{F}_{2^2} \longrightarrow \mathbb{F}_2[\mathbf{U}]$ *such that* $\psi(\alpha)=\mathbf{U}$ *and* $\psi(0)=\mathbf{O}$*, which is a field isomorphism.*

It is possible to construct the ring isomorphism $\Psi : \mathrm{Mat}_{r\times r}(\mathbb{F}_{2^2}) \longrightarrow \mathrm{Mat}_{r\times r}(\mathbb{F}_2[\mathbf{U}])$*, such that* $\Psi(\mathbf{A})=[\psi(a_{ij})]$*, for* $\mathbf{A}=[a_{ij}] \in \mathrm{Mat}_{r\times r}(\mathbb{F}_{2^2})$*.*

Now, consider the primitive polynomial $V(\mathsf{x}) = \alpha + \alpha^2\mathsf{x} + \mathsf{x}^2 + \mathsf{x}^3 \in \mathbb{F}_{2^2}[\mathsf{x}]$ *whose companion matrix is*

$$\mathbf{V} = \begin{bmatrix} 0 & 0 & \alpha \\ 1 & 0 & \alpha^2 \\ 0 & 1 & 1 \end{bmatrix}.$$

Then,

$$\Psi(\mathbf{V}) = \begin{bmatrix} \mathbf{O} & \mathbf{O} & \mathbf{U} \\ \mathbf{I} & \mathbf{O} & \mathbf{U}^2 \\ \mathbf{O} & \mathbf{I} & \mathbf{I} \end{bmatrix} = \left[\begin{array}{cc:cc:cc} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ \hdashline 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ \hdashline 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right]$$

and by Theorem 1 the characteristic polynomial

$$C(\mathsf{x}) = \det\left(\mathsf{x}\mathbf{I} - \Psi(\mathbf{V})\right) = 1 + \mathsf{x} + \mathsf{x}^6$$

of matrix $\Psi(\mathbf{V})$ *is a primitive polynomial with coefficients in* $\mathbb{F}_2$ *and degree* 6*.*

Summarizing, with a polynomial of degree 3 *over* $\mathbb{F}_{2^2}[\mathsf{x}]$ *and a polynomial of degree* 2 *over* $\mathbb{F}_2[\mathsf{x}]$ *we found a polynomial* $C(\mathsf{x})$ *of degree* 6 *over* $\mathbb{F}_2[\mathsf{x}]$*.*

Acknowledgements

The work of the first author was supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) with number of process $2015/07246-0$. The second author was partially supported by grant MIMECO MTM2015-69138-REDT.

References

[1] Jacob T. B. Beard, Jr., and Karen I. West. Some primitive polynomials of the third kind. *Mathematics of Computation*, 28(128):1166–1167, 1974.

[2] Sara D. Cardell, Joan-Josep Climent, and Verónica Requena. A construction of MDS array codes. *WIT Transactions on Information and Communication Technologies*, 45:47–58, 2013.

[3] Solomon W. Golomb. *Shift Register-Sequences*. Aegean Park Press, Laguna Hill, California, 1982.

[4] Tom Hansen and Gary L. Mullen. Primitive polynomials over finite fields. 59(200):639–643, 1992.

[5] Rudolf Lidl and Harald Niederreiter. *Introduction to Finite Fields and Their Applications*. Cambridge University Press, New York, NY, 1986.

[6] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone. *Handbook of Applied Cryptography*. CRC Press, Boca Raton, FL, 1996.

[7] Wayne Stahnke. Primitive binary polynomials. 27(124):977–980, 1973.

[8] E. J. Watson. Primitive polynomials (mod 2). 16:368–269, 1962.

[9] Neal Zierler and John Brillhart. On primitive trinomials (mod 2). *Information and control*, 13(6):541–554, 1968.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The modified self-shrinking generator via the generalized self-shrinking generator

Sara D. Cardell[1] **and Amparo Fúster-Sabater**[2]

[1] *Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas*

[2] *Instituto de Tecnologías Físicas y de la Información, CSIC*

emails: sdcardell@ime.unicamp.br, amparo@iec.csic.es

Abstract

The modified self-shrinking generator was recently designed for stream cipher applications. This cryptographic keystream generator is a new and improved version of the self-shrinking generator. However, it is possible to see that the sequences produced by both generators are also obtained as output sequences of the generalized self-shrinking generator.

Key words: modified self-shrinking generator, generalized self-shrinking generator, characteristic polynomial.

The **modified self-shrinking generator** (MSSG), introduced by Kanso in 2010 [3], is a special case of the self-shrinking generator [4], where the PN-sequence $\{u_i\}$ generated by a maximum-length LFSR [1] is self-decimated. Here the decimation rule is very simple and can be described as follows: Given three consecutive bits $\{u_{2i}, u_{2i+1}, u_{2i+2}\}$, $i = 0, 1, 2, \ldots$ the output sequence $\{s_j\}$, known as **modified self-shrunken sequence**, is computed as

$$\begin{cases} \text{If } u_{2i} + u_{2i+1} = 1 \text{ then } s_j = u_{2i+2} \\ \text{If } u_{2i} + u_{2i+1} = 0 \text{ then } u_{2i+2} \text{ is discarded.} \end{cases}$$

If L (odd) is the length of the maximum-length LFSR that generates $\{u_i\}$, then the linear complexity LC of the corresponding modified self-shrunken sequence satisfies:

$$2^{\lfloor \frac{L}{3} \rfloor - 1} \leq LC \leq 2^{L-1} - (L-2),$$

and the period T of the sequence satisfies:

$$2^{\lfloor \frac{L}{3} \rfloor} \leq T \leq 2^{L-1}$$

as proved in [3]. As usual, the key of this generator is the initial state of the register that generates the PN-sequence $\{u_i\}$.

Example 1 *Consider the LFSR of length $L = 3$ with characteristic polynomial $p(x) = 1 + x^2 + x^3$ and initial state $IS = \{1\ 0\ 0\}$. The PN-sequence generated is* $1001110\ldots$ *with period $T = 2^3 - 1$.*

The modified self-shrunken sequence in this case is given by: $\{0011\ldots\}$ *and has period $T = 4$ and characteristic polynomial $(1 + x)^3$. Thus, the linear complexity of this modified self-shrunken sequence is $LC = 3$.* ■

Let $\{a_i\}$ be PN-sequence generated by a maximum-length LFSR of L stages. Let G be and L-dimensional binary vector $G = (g_0, g_1, g_2, ..., g_{L-1}) \in \mathbb{F}_2^L$ and $\{v_i\}$ a sequence defined as: $v_i = g_0 a_i + g_1 a_{i-1} + g_2 a_{i-2} + \cdots + g_{L-1} a_{i-L+1} \mod(2^L - 1)$.

For $n \geq 0$, the decimation rule is given by:

$$\begin{cases} \text{If } a_i = 1 \text{ then } b_j = v_i \\ \text{If } a_i = 0 \text{ then } v_i \text{ is discarded.} \end{cases}$$

The output sequence $\{b_j\}$,denoted by $b(G)$, is called **generalized self-shrunken sequence** associated with G (see [2]).

When G ranges over $\mathbb{F}_2^L$, $\{v_i\}$ corresponds to the $2^L - 1$ possible shifts of $\{a_i\}$. Furthermore, the set of sequences denoted by $B(a) = \{b(G) \mid G \in \mathbb{F}_2^L\}$ is the family of generalized self-shrunken sequences based on the PN-sequence $\{a_i\}$.

Example 2 *For an LFSR of length* 3 *whose characteristic polynomial is $p'(x) = 1 + x + x^3$ and output PN-sequence* $\{1\ 1\ 1\ 0\ 0\ 1\ 0\}$*, we get the generalized self-shrunken sequences shown in the following table:*

G	v	$b(G)$
0 0 0	$\underline{0}\ \underline{0}\ \underline{0}\ 0\ 0\ \underline{0}\ 0$	0 0 0 0
0 0 1	$\underline{1}\ \underline{0}\ \underline{1}\ 1\ 1\ \underline{0}\ 0$	1 0 1 0
0 1 0	$\underline{0}\ \underline{1}\ \underline{1}\ 1\ 0\ \underline{0}\ 1$	0 1 1 0
0 1 1	$\underline{1}\ \underline{1}\ \underline{0}\ 0\ 1\ \underline{0}\ 1$	1 1 0 0
1 0 0	$\underline{1}\ \underline{1}\ \underline{1}\ 1\ 1\ \underline{1}\ 1$	1 1 1 1
1 0 1	$\underline{0}\ \underline{1}\ \underline{0}\ 1\ 1\ \underline{1}\ 0$	0 1 0 1
1 1 0	$\underline{1}\ \underline{0}\ \underline{0}\ 1\ 0\ \underline{1}\ 1$	1 0 0 1
1 1 1	$\underline{0}\ \underline{0}\ \underline{1}\ 0\ 1\ \underline{1}\ 1$	0 0 1 1
	1 1 1 0 0 1 0	

 ISBN: 978-84-608-6082-2

The modified self-shrunken sequence computed in Example 1 with polynomial $p(x) = 1 + x^2 + x^3$*, can be also obtained through the generalized self-shrunken sequence with polynomial* $p'(x) = 1 + x + x^3$ *(see the last sequence in the previous table).*

In general, the modified self-shrunken sequence obtained with a primitive polynomial $p(x) \in \mathbb{F}_2[x]$ of degree L can be obtained as well through the generalized self-shrunken sequence with a primitive polynomial $p'(x) \in \mathbb{F}_2[x]$ of the same degree. This polynomial corresponds to:

$$p(x) = (x + \alpha^3)(x + \alpha^6)(x + \alpha^{12}) \cdots (x + \alpha^{3 \cdot 2^{L-1}}),$$

The implementation of the modified self-shrunken sequence via the generalized self-shrinking sequence is more efficient, since we need a fewer number of computations and a small quantity of bits involved in these calculations.

Acknowledgements

The work of the first author was supported by FAPESP with number of process 2015/07246-0. The work of the second author was supported by both Ministerio de Economía y Competitividad, Spain, under grant number TIN2014-55325-C2-1-R (ProCriCiS), and Comunidad de Madrid, Spain, under grant number S2013/ICE-3095-CM (CIBERDINE).

References

[1] Solomon W. Golomb. *Shift Register-Sequences*. Aegean Park Press, Laguna Hill, California, 1982.

[2] Yupu Hu and Guozhen Xiao. Generalized self-shrinking generator. 50(4):714–719, 2004.

[3] Ali Kanso. Modified self-shrinking generator. *Computers and Electrical Engineering*, 36(1):993–1001, 2010.

[4] Willi Meier and Othmar Staffelbach. The self-shrinking generator. In Christian Cachin and Jan Camenisch, editors, *Advances in Cryptology – EUROCRYPT 1994*, volume 950, pages 205–214. Springer-Verlag, 1994.

Volume II

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Minimal Faithful Upper-Triangular Matrix Representations for Solvable Lie Algebras

Manuel Ceballos[1], Juan Núñez[2] and Ángel F. Tenorio[3]

[1] *Departamento de Ingeniería, Universidad Loyola Andalucía*

[2] *Departamento de Geometría y Topología, Facultad de Matemáticas. Universidad de Sevilla.*

[3] *Dpto. de Economía, Métodos Cuantitativos e Historia Económica, Escuela Politécnica Superior. Universidad Pablo de Olavide.*

emails: mceballos@uloyola.es, jnvaldes@us.es, aftenorio@upo.es

Abstract

Every finite-dimensional complex solvable Lie algebra can be represented as a matrix Lie algebra, with upper-triangular square matrices as elements. Nevertheless, the minimal order of these matrices is unknown in general. In this paper, we draft a method to compute both that minimal order and a matrix representation of such an order for a given solvable Lie algebra. As application of this procedure, we compute minimal faithful matrix representations for several families of Lie algebras with arbitrary dimensions.

Key words: solvable Lie algebra, faithful matrix representation, minimal representation, symbolic computation, non-numerical algorithm
MSC 2000: 17B30, 17B05, 17–08, 68W30, 68W05.

1 Introduction

In virtue of Ado's Theorem, every finite-dimensional complex Lie algebra is isomorphic to a Lie subalgebra of the general linear algebra $\mathfrak{gl}(n;\mathbb{C})$ of complex $n \times n$ matrices, for some $n \in \mathbb{N}$. This paper deals with finding Lie subalgebras of Lie algebra $\mathfrak{h}_n$, of $n \times n$ upper-triangular matrices, being isomorphic to solvable Lie algebras; since every finite-dimensional solvable Lie algebra is isomorphic to a subalgebra of $\mathfrak{h}_n$, for some $n \in \mathbb{N}$, according to [6, Proposition 3.7.3]. Consequently, it is interesting to determine which is the minimal $n \in \mathbb{N}$

such that a given Lie algebra $\mathfrak{g}$ is isomorphic to some Lie subalgebra of $\mathfrak{h}_n$; i.e. finding out the minimal dimension of faithful representations of $\mathfrak{g}$ by using $n \times n$ upper-triangular matrices.

This topic has been previously studied in the literature. For example, Benjumea et al. [1] designed an algorithmic procedure which computed the minimal faithful unitriangular matrix representations of nilpotent Lie algebras. The complete list of these representations were given by Benjumea et al. [2] for nilpotent Lie algebras of dimension less than 6. With respect to filiform Lie algebras, their representations were studied in [4] for dimension less than 9. Our main goal is to obtain tools which ease the advance for the above-mentioned research. To do this, we introduce the design of an algorithm (which has been implemented) to compute minimal faithful matrix representations of solvable Lie algebras.

2 Preliminaries

From here on, we have only considered finite-dimensional Lie algebras over the complex number field $\mathbb{C}$. The reader can consult [6] for a comprehensive review on Lie algebras.

Given a Lie algebra $\mathfrak{g}$, its *derived series* is defined as follows

$$\mathcal{C}_1(\mathfrak{g}) = \mathfrak{g},\ \mathcal{C}_2(\mathfrak{g}) = [\mathfrak{g},\mathfrak{g}],\ \mathcal{C}_3(\mathfrak{g}) = [\mathcal{C}_2(\mathfrak{g}),\mathcal{C}_2(\mathfrak{g})],\ \ldots,\ \mathcal{C}_k(\mathfrak{g}) = [\mathcal{C}_{k-1}(\mathfrak{g}),\mathcal{C}_{k-1}(\mathfrak{g})],\ \ldots$$

If there exists a natural integer m such that $\mathcal{C}_m(\mathfrak{g}) \equiv 0$, then $\mathfrak{g}$ is said to be *solvable*. We have the following

Proposition 1

If $\mathfrak{h}$ is a Lie subalgebra of a given Lie algebra $\mathfrak{g}$, then $\mathcal{C}_k(\mathfrak{h}) \subseteq \mathcal{C}_k(\mathfrak{g})$, for all $k \in \mathbb{N}$.

Given $n \in \mathbb{N}$, the complex solvable Lie algebra $\mathfrak{h}_n$ consists of $n \times n$ upper-triangular matrices; i.e. its elements can be expressed as

$$h_n(x_{r,s}) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ 0 & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & x_{nn} \end{pmatrix}, \qquad \text{with } x_{r,s} \in \mathbb{C},\ \text{for } 1 \leq r \leq s \leq n.$$

The dimension of $\mathfrak{h}_n$ is $\frac{n(n+1)}{2}$ and the nonzero brackets are

$$\begin{array}{ll} [X_{i,j}, X_{j,k}] = X_{i,k}, & \forall\ 1 \leq i < j < k \leq n; \\ [X_{i,i}, X_{i,j}] = X_{i,j}, & 1 \leq i < j \leq n; \\ [X_{k,i}, X_{i,i}] = X_{k,i}, & \forall\ k \leq i \leq n. \end{array}$$

with respect to the basis $\mathcal{B}_n = \{X_{i,j} = h_n(x_{r,s}) \,|\, x_{r,s} = \delta_{r,i} \cdot \delta_{s,j},\ \text{for } 1 \leq r \leq s \leq n\}_{1 \leq i \leq j \leq n}$, where δ denotes the Krönecker delta function.

 ISBN: 978-84-608-6082-2

3 Minimal Matrix Representations

Given a Lie algebra $\mathfrak{g}$, a *representation* of $\mathfrak{g}$ in $\mathbb{C}^n$ is a Lie-algebra homomorphism $\phi : \mathfrak{g} \to \mathfrak{gl}(\mathbb{C}, n)$. The *dimension* of this representation is the value of $n \in \mathbb{N}$. Ado's theorem assures the existence of a linear injective representation (i.e. a *faithful representation*) on a finite-dimensional vector space for every finite-dimensional Lie algebra over a field of characteristic zero.

Representations are usually defined as $\mathfrak{g}$-*modules*; i.e. Lie-algebra homomorphisms from $\mathfrak{g}$ to Lie algebra $\mathfrak{gl}(V)$ of endomorphisms over an arbitrary n-dimensional vector space V (see [5]).

With respect to minimal representations of Lie algebras, Burde [3] introduced the invariant $\mu(\mathfrak{g})$ for any given Lie algebra $\mathfrak{g}$

$$\mu(\mathfrak{g}) = \min\{\dim(M) \mid M \text{ is a faithful } \mathfrak{g}\text{-module}\}.$$

The goal of this paper is to study matrix faithful representations of solvable Lie algebras. More concretely, we are looking for minimal faithful matrix representations being contained in $\mathfrak{h}_m$, for some $m \in \mathbb{N}$. Indeed, given a solvable Lie algebra $\mathfrak{g}$, our goal is to determine the minimal value m such that $\mathfrak{g}$ is isomorphic to a Lie subalgebra of $\mathfrak{h}_m$, but not of $\mathfrak{h}_{m-1}$. This value is also an invariant of $\mathfrak{g}$ and can be expressed as

$$\bar{\mu}(\mathfrak{g}) = \min\{m \in \mathbb{N} \mid \exists \text{ subalgebra of } \mathfrak{h}_m \text{ isomorphic to } \mathfrak{g}\}.$$

In general, invariants $\mu(\mathfrak{g})$ and $\bar{\mu}(\mathfrak{g})$ can be different from each other.

To carry out the computation of a minimal faithful matrix representations for a given n-dimensional solvable Lie algebra $\mathfrak{g}$ by using Lie algebras $\mathfrak{h}_m$, we have sketched an algorithmic method which only requires the law of $\mathfrak{g}$ as input. The steps of this algorithm are the following

1. Construct the derived series of $\mathfrak{g}$ and look for the first natural integer k such that this series fits in with that associated with $\mathfrak{h}_k$. Consequently, we are interested in the first k verifying $\mathcal{C}_i(\mathfrak{g}) \subseteq \mathcal{C}_i(\mathfrak{h}_k)$, for all $i \in \mathbb{N}$. This is in virtue of Proposition 1.

2. Express the vectors in the basis $\{e_i\}_{i=1}^n$ of $\mathfrak{g}$ as linear combinations of basis $\mathcal{B}_k$ of $\mathfrak{h}_k$ (i.e. express $\mathfrak{g}$ as a subalgebra of $\mathfrak{h}_k$):

$$e_h = \sum_{1 \le i \le j \le k} \lambda_{i,j}^h X_{i,j}, \quad \text{for } 1 \le h \le n,$$

   taking into account the possible simplifications resulting from Proposition 1.

3. Compute bracket $[e_i, e_j]$ for $1 \le i \le j \le n$ and compare coordinate to coordinate with the results in the laws of $\mathfrak{g}$, but expressed in terms of the basis $\mathcal{B}_k$ of $\mathfrak{h}_k$.

 ISBN: 978-84-608-6082-2

4. Solve the previous system and consider one of the solutions corresponding to a set of linearly independent vectors (i.e. this solution is the faithful matrix representation searched). If there are no such solutions, then go back to Step 2 and repeat each step with Lie algebra $\mathfrak{h}_{k+1}$ since Lie algebra $\mathfrak{g}$ cannot be represented as a Lie subalgebra of $\mathfrak{h}_k$.

Since we start with $k = 1$ and k increases one unit when there are no representations in $\mathfrak{h}_k$, we can assert that the representation obtained in the previous algorithm is minimal.

4 Some tips for a successful implementation

According to the design previously sketched, the implementation of our procedure should be organized to give answer to the following issues in this order

- Computing the law of solvable Lie algebra $\mathfrak{h}_n$.
- Computing two lists to save the dimension of the ideals in the derived series of $\mathfrak{h}_n$ and $\mathfrak{g}$.
- Determining the minimal dimension k such that the derived series of $\mathfrak{g}$ is compatible with that of $\mathfrak{h}_k$ (i.e. the smallest k which should be considered to find a faithful matrix representation of $\mathfrak{g}$).
- Expressing all the vectors in the basis of $\mathfrak{g}$ as linear combinations of basis $\mathcal{B}_k$ of $\mathfrak{h}_k$.
- Imposing the law of both Lie algebras and applying Proposition 1 to obtain an equation system.
- With a symbolic computation package, solving the system of equations resulting from the previous expressions. Since equations are polynomial, every symbolic computation package computes efficiently the algebraic expression of the set of solutions.
- Checking if some of the solutions corresponds to a list of linearly independent vectors.
- If there exists such a solution, then some free coefficients may appear because the faithful matrix representation of $\mathfrak{g}$ may be not unique. In order to obtain a simpler representative, we consider as many null coefficients as possible.

When implementing this algorithm (in our case with Maple), we can compute a minimal faithful matrix representation of any solvable Lie algebra and, more concretely, the list of representations for each dimension whose classification is known.

 ISBN: 978-84-608-6082-2

Acknowledgment

This work has been partially supported by MTM2013-40455-P.

References

[1] J. C. Benjumea, F. J. Echarte, J. Núñez and A. F. Tenorio, *A method to obtain the Lie group associated with a nilpotent Lie algebra*, Comput. Math. Appl. 51 (2006) 1493–1506.

[2] J. C. Benjumea, J. Núñez and A. F. Tenorio, *Minimal linear representations of the low-dimensional nilpotent Lie algebras*, *Math. Scand.* 102 (2008) 17–26.

[3] D. Burde, *On a refinement of Ado's Theorem*, Arch. Math. (Basel) 70 (1998) 118–127.

[4] M. Ceballos, J. Núñez and A. F. Tenorio, *Representing Filiform Lie Algebras Minimally and Faithfully by Strictly Upper-Triangular Matrices*, J. Algebra Appl. 12 (2013), 1250196, 15pp.

[5] W. Fulton and J. Harris, *Representation theory: a first course*, Springer-Verlag, 1991.

[6] V. S. Varadarajan, *Lie Groups, Lie Algebras and their Representations*. Selected Monographies 17, Collæge Press, Beijing, 1998.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Two-Stage Intra Prediction Algorithm for HEVC

Gabriel Cebrián-Márquez[1], José Luis Martínez[1] and Pedro Cuenca[1]

[1] *High-Performance Networks and Architectures (RAAP), Albacete Research Institute of Informatics (I3A), University of Castilla-La Mancha, Albacete, Spain*

emails: Gabriel.Cebrian@uclm.es, JoseLuis.Martinez@uclm.es, Pedro.Cuenca@uclm.es

Abstract

The unceasing demands for high quality multimedia contents and the advent of new resolutions such as Ultra High Definition (UHD) motivated the development of the High Efficiency Video Coding (HEVC) standard, which outperforms prior standards by up to 50% in terms of coding efficiency. While this improvement meets the aforementioned demands, it also involves higher computational complexities in the encoder. For this reason, fast and efficient coding algorithms are now a requirement of HEVC-compliant real-time encoders. In this regard, this paper proposes a pre-analysis algorithm designed to provide coding information to the intra module of the encoding stage. As a result, experiments show that the algorithm is able to reduce the encoding time by up to 9.15% at the expense of negligible losses in terms of coding efficiency.

Key words: HEVC, H.265, Pre-Analysis, Intra Coding

1 Introduction

For more than ten years, H.264/Advanced Video Coding (AVC) [1] has established itself as the most widespread video compression standard for all types of applications and scenarios, e.g. Blu-ray and also various High-Definition (HD) television broadcasts. However, the advent of new video formats such as the Ultra High Definition (UHD) resolution and the ensuing increase in bit rate along with the demands for higher video quality motivated the development of the High Efficiency Video Coding (HEVC) standard [2]. Defined by the Joint Collaborative Team on Video Coding (JCT-VC) in early 2013, HEVC roughly doubles the rate-distortion (R-D) performance of H.264/AVC, which implies nearly 50% bit

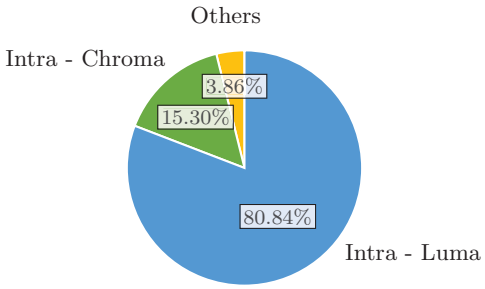


Figure 1: Time profile of the Kimono sequence using All Intra configuration and $QP = 22$

rate reduction for the same video quality [3]. This improvement in coding efficiency comes, however, at the expense of an extremely high computational complexity [4].

In order to reduce the processing time, HEVC introduces some high-level tools, such as tiles and wavefronts, with the aim of allowing the parallel encoding and decoding of a video sequence. These tools are, however, insufficient to achieve real-time video compression. For this reason, the research community is making a considerable effort to find complementary ways of reducing the computational complexity of the encoder not only from the point of view of parallelism, but also algorithmically.

Figure 1 shows the profiling results obtained from the baseline HEVC encoder for the *Kimono* 1080p sequence, using the All Intra configuration and setting the quantization parameter (QP) to 32. As can be seen, the vast majority of the time is devoted to the intra prediction module of the encoder, becoming the most computationally expensive operation in the encoder for this mode, whereas only 3.86% of the total execution time corresponds to the remaining encoder modules (labelled as "Others"), including transform, quantization, entropy coding, and in-loop filtering. The complexity of the intra module can be justified by the large number of modes it has to test for every prediction unit (PU), namely 33 directional modes, and the DC and planar modes.

Most multimedia applications require real-time encoders, especially those which make use of the All Intra coding pattern. For this reason, these encoders typically implement fast algorithms and techniques that support the encoding in several ways. In particular, many commercial encoders include a so-called pre-analysis algorithm, which is responsible for providing the encoder some preliminary information about the input pictures and the encoding process. This information is used by the encoder for different purposes, e.g. reducing the encoding time or deciding the encoding pattern in Random Access configurations. In this paper, we design and implement a pre-analysis algorithm for HEVC with the aim of reducing the computational complexity of the intra coding. In this way, the aforementioned algorithm performs a fast calculation of the costs of the different 35 intra modes, which is

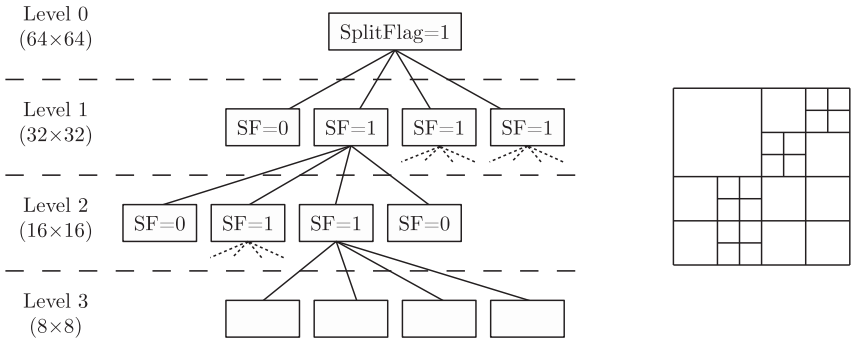


Figure 2: Quadtree partitioning structure in HEVC

later used in the intra module of the encoder. As a result, the encoding time is reduced having little impact on the coding efficiency. It also has to be noted that this algorithm can be applied to several other video codecs with complex partitioning schemes as in HEVC.

The rest of this paper is organized as follows. Section 2 presents briefly the principles of HEVC. Section 3 covers some of the most relevant related works in the topic. Following this, Sections 4 and 5 provide an overview of the pre-analysis architecture and the proposed algorithm, respectively. An experimental evaluation of this algorithm is carried out in Section 6, showing its results in terms of time reduction and coding efficiency. Finally, in Section 7 conclusions are drawn.

2 Technical Background

Not only did HEVC introduce new coding tools in the standard, but also it improved others which were present in H.264/AVC, notably increasing the overall compression efficiency. In fact, one of the most important changes affects the picture partitioning. HEVC discards the term macroblock (MB) and replaces it with a new structure: the coding tree unit (CTU). CTUs are squared regions into which the picture is divided. Each CTU, whose size is typically 64×64, can be recursively partitioned into four coding units (CU). These CUs, ranging from 8×8 to 64×64 pixels, are composed of PUs and transform units (TU). CTUs and CUs form a quadtree structure as the one shown in Figure 2.

With regard to the partitioning of CUs into PUs, HEVC defines up to eight possible inter and intra partitions for each CU size (2N×2N, 2N×N, N×2N, N×N, 2N×nU, 2N×nD, nL×2N and nR×2N) to determine the optimal trade-off between rate and distortion. The last four PU types correspond to the asymmetric motion partitioning (AMP), and they were not present in previous standards. The introduction of these new sizes involves a large increase in the complexity of the encoder, but it also allows the standard to be more flexible

to adapt itself for edges, specially with larger CU sizes. Other new features in HEVC include a total of 35 different intra coding modes, the residual quadtree (RQT) in the case of the transform, or the new in-loop filters.

3 Related Work

Even though the standard already defines parallel tools that assist the encoding process, several related works focus on further exploiting the parallelism of HEVC. For example, authors in [5] describe the Overlapped Wavefront (OWF) algorithm, an improved version of the Wavefront Parallel Processing (WPP) algorithm defined in the standard that solves the problem of the so-called ramping inefficiencies at the beginning and at the end of the frames. Alternatively, an analysis of the existing dependencies in the motion estimation (ME) module is shown in [6], where a highly parallel framework designed for many-core processors is proposed. Many other works make use of heterogeneous platforms, such as those based on GPU [7]. However, as mentioned in the introduction of this paper, these techniques are not enough to overcome the computational complexity of the encoder, and fast encoding algorithms are still necessary.

Other related works focus on reducing the computational complexity of the encoder from an algorithmic point of view. The approach followed by many is to prune the decision tree in order to skip those partitions in which it is unlikely to find a better prediction candidate than the current one. For example, in the case of intra coding, authors in [8] make use of Bayesian techniques to determine the splitting or pruning of the CTU tree based on a set of statistical parameters, while authors in [9] show that the splitting of the current CU can be determined by the existing relationship in terms of mode and cost with the neighbouring CUs. All of these techniques perform their operations on a higher level, and thus they are complementary to the algorithm proposed in this paper. Some other works focus on reducing the complexity of the intra mode selection such as [10, 11]. However, while the authors obtain a considerable speed-up, the resulting BD-rate is higher than the one obtained in this paper.

With regard to pre-analysis algorithms, it is worth mentioning the look-ahead module of x264 H.264/AVC encoder [12, 13]. This module splits a subsampled version of the input frame into fixed-size blocks, and calculates their corresponding intra and inter costs. Some of the ways in which this information is used in the encoder include determining the encoding pattern, speeding up some operations and adjusting the parameters of the rate-control module. However, this technique is targeted for H.264/AVC, which involves less block partitions than HEVC, and even though it has been assimilated in x265 [14] (the version of x264 for HEVC), it does not completely meet the requirements of the new standard. Other uses for the pre-analysis algorithm range from the adjustment of the rate-control algorithm [15] to the analysis of the texture of a picture for CU-size selection in intra

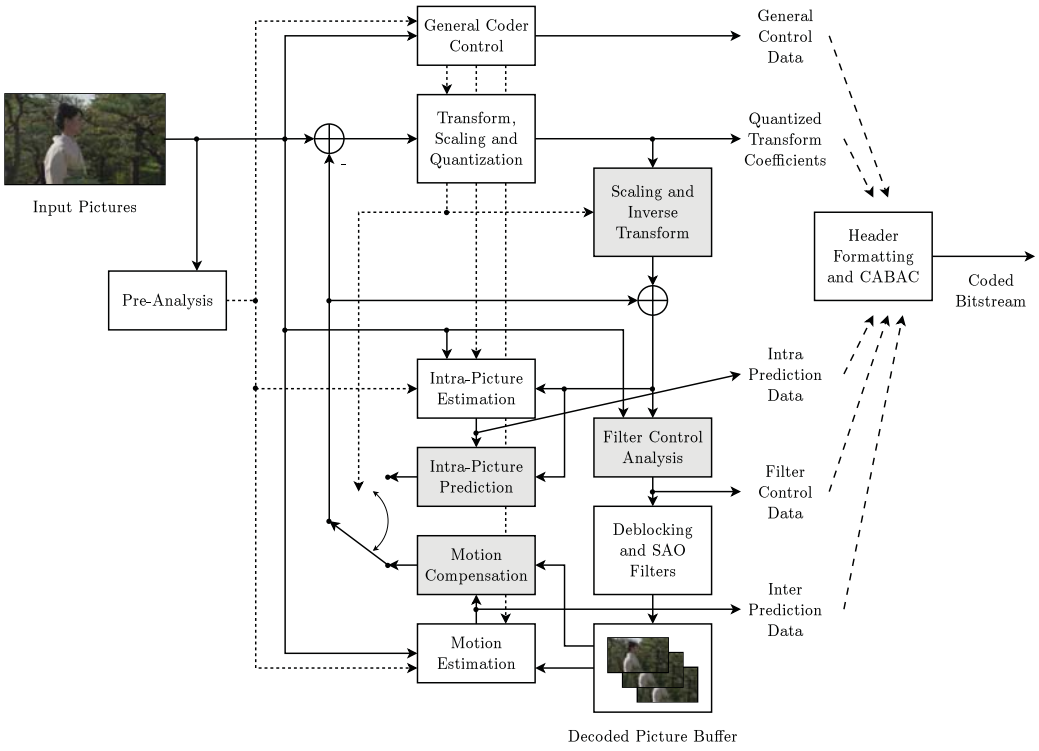


Figure 3: Typical HEVC encoder architecture with a pre-analysis stage (shared elements with the decoder are shaded in gray)

configurations [16], and are completely complementary to the algorithm proposed in this paper.

4 Pre-Analysis Algorithm Architecture

As introduced in the first section of this paper, a pre-analysis algorithm is responsible for obtaining some preliminary information from the input frames, which is in turn used by the encoder in various ways. Figure 3 shows the architecture of a typical HEVC-compliant encoder, in which a pre-analysis stage has been added. The decoder processing blocks that the encoder duplicates have been shaded in gray. As can be seen, the input pictures are analysed by this stage before any encoding is performed. This involves that any information has to be gathered from the original samples, i.e. without any partitioning or neighbour predictors. In spite of this, the amount of information that can be extracted and predicted by the pre-analysis algorithm is endless, including motion information, intra modes, prediction

costs, coding pattern, partitioning, etc. This information is thus used in the many modules that constitute the encoder to control or aid the encoding process.

In prior works, we focused our efforts on implementing a pre-analysis algorithm for inter prediction. The obtained motion information can be used to reduce the number of positions checked in the search area [17, 18] or to limit the number of reference frames [18]. As a result, the total encoding time is notably reduced at the expense of negligible losses in terms of coding efficiency. Nevertheless, this pre-analysis algorithm is only valid for those coding configurations that make use of inter prediction, such as Random Access. For this reason, in this paper we design and develop an alternative but complementary algorithm for the intra prediction module, especially targeted to All Intra configurations.

5 Proposed Pre-Analysis Algorithm for Intra Prediction

The aim of the pre-analysis stage is to estimate the best intra mode and its corresponding cost for each PU in a fast way. This is achieved by carrying out a similar process to that of the intra module. The latter, however, involves large computation times, as it needs to test every possible mode and perform the transform operation. For this reason, this operation is simplified in the pre-analysis stage by using less complex operations.

One way in which the intra operation can be performed in the pre-analysis stage is by dividing the input frames into fixed-size block partitions, resulting in a grid of square blocks. For each of these blocks, the best intra mode is calculated (directional, DC or planar), which in turn provides a prediction cost. However, we proved experimentally that extrapolating the best intra mode from the pre-analysis data for every PU is not really possible if only one block size is utilised, given the large range of partitioning possibilities that HEVC offers. For this reason, our proposal divides the input frames into fixed-size block partitions that vary from 4×4 to 64×64, including 8×8, 16×16 and 32×32. In this way, this preliminary information can be more easily adapted to the different PUs the encoder might test. In fact, for each PU there is a pre-analysis block that exactly matches in position and size.

In order to obtain the most reliable prediction possible, all the 33 directional modes are checked for every block, along with the DC and planar modes. With the aim of performing a fast estimation, the sum of absolute differences (SAD) measurement is used instead of the Hadamard transform. While the latter offers a better representation of the distortion caused by the prediction, the former also allows to compare several prediction modes, but involving a notably lower computational complexity.

It also has to be taken into account that a pre-analysis stage has some limitations compared to the encoding stage itself. Nevertheless, it is possible to address them so that the results obtained in the former are virtually comparable to those of the latter. As a result, the effects on the resulting coding efficiency are negligible. Some of these limitations are:

- As mentioned in Section 4, the pre-analysis stage performs its operations on the original input frames. The encoder, however, makes use of the neighbouring reconstructed samples for the intra prediction. Nevertheless, considering that the reconstructed samples are fairly similar to the original ones, especially for high QP values, it is reasonable to assume that the prediction in both cases will be similar enough to determine which prediction modes are best.

- By the same token, there is no neighbour information other than the one that this stage itself calculates. As a result, it is not possible to know which samples might be or might not be available for prediction at the encoding stage. However, the standard defines some sample substitution techniques that ensures that there is always present the required neighbouring samples to perform intra prediction [3]. In this way, it is possible to assume that in many cases the original samples will be similar to those that the encoder will create in the substitution process, and thus this is the approach that the proposed pre-analysis algorithm follows.

In addition to these two considerations, it also has to be taken into account that the encoder makes use of an R-D model such as:

$$J = Distortion_{intra} + \lambda \cdot Rate_{intra} \tag{1}$$

where J represents the cost function, $Distortion_{intra}$ the distortion function (typically Hadamard), λ is the Lagrangian multiplier which depends on the QP, and $Rate_{intra}$ represents the bits required to code the prediction mode. Nevertheless, while it is possible to calculate the λ value and estimate the $Distortion_{intra}$ in the pre-analysis stage, it is not feasible to predict the value of $Rate_{intra}$, given that the neighbouring blocks have not been entropy-coded at that moment. For this reason, only the distortion is estimated in this stage, postponing the R-D calculation until the moment in which the encoder performs intra prediction.

With regard to the integration of this algorithm, it has to be noted that the pre-analysis stage can be part of any HEVC-compliant encoder. However, in order to show how these encoders can benefit from this algorithm, the remainder of this section will detail its integration into the HEVC Test Model (HM) reference software [19]. This encoder performs a 3-step intra prediction algorithm in order to obtain the best intra mode for a given PU. In the first step, a rough prediction is carried out, in which all the 35 intra modes are tested. In the second step, a set of the best candidates found in the previous step is considered along with the most probable modes (MPM), performing a more complex prediction and a transform of the resulting residual. Finally, the last step consists on choosing the best candidate out of the ones from the second step and performing the full RQT partitioning, which results in the definitive R-D cost for the corresponding intra mode.

As can be seen, the integration of our proposal in this case is fairly immediate. As the proposed pre-analysis algorithm already estimates the prediction cost of the 35 intra modes, the first step of the encoding stage in HM can be skipped. Therefore, the set of best candidates for the second step would be simply obtained from the pre-analysis scores. While it might seem there is no time reduction from this, actually the operations performed in the pre-analysis stage are much less complex than the ones in the encoding stage. For example, there is a lower number of conditions to check, and also the distortion algorithm being used is SAD and not Hadamard, as mentioned before. The next section will show in which ways the encoder is influenced by the proposed pre-analysis algorithm.

6 Performance Evaluation

The experiments have been executed on the HM 16.6 reference software, following the guidelines and coding conditions provided by the document elaborated by the JCT-VC [20]. In this regard, All Intra has been the selected configuration, as it is the configuration to which our proposal is targeted. The QP values being tested are 22, 27, 32 and 37. The encodings have been carried out using the Main profile and 8-bit depth. Sequences of classes A to D according to the JCT-VC classification have been used, which include the following resolutions: 2560×1600 (A), 1920×1080 (B), 832×480 (C), and 416×240 (D). The rest of the configuration parameters have all been kept to their default values.

The hardware platform used in the experiments is composed of an Intel® Xeon® E5-2630L v3 CPU running at 1.80 GHz and 16 GB of main memory. The encoder has been compiled with GCC 4.8.5-4 and executed on CentOS 7 (Linux 3.10.0-327). Turbo Boost has been disabled to achieve the reproducibility of the results.

The results will be provided in terms of time reduction (TR) and Bjøntegaard delta rate (BD-rate). The BD-rate is a measure of coding efficiency that represents the percentage of bit rate variation between two encodings with the same objective quality [21].

Table 1 shows the results obtained by the encoder when the pre-analysis algorithm is enabled. As can be seen, making use of the information gathered by the pre-analysis stage, the encoder is able to save up to 9.15% of the total encoding time (7.07% on average). As mentioned before, these time savings are derived from the fact that the pre-analysis stage performs less complex operations compared to the baseline encoder, more specifically because of the substitution of the Hadamard algorithm with the SAD operation. But even if Hadamard was used in the pre-analysis stage, it would still be possible to obtain time savings, which demonstrates the fact that this stage is less computationally complex, as it needs to check less conditions than the encoder.

This table also shows, however, that the pre-analysis algorithm has an effect on the coding efficiency. In the experiments, the obtained average BD-rate is 0.46%, which is considered negligible for most multimedia applications. This means that the bit rate is

Table 1: Results of the proposed algorithm for the All Intra configuration

Class	Video sequence	BD-rate (%)	Encoding TR (%)
A	Traffic	0.51	7.96
	PeopleOnStreet	0.59	7.87
	NebutaFestival	0.20	6.23
	SteamLocomotive	0.43	8.62
B	Kimono	0.53	9.15
	ParkScene	0.40	7.31
	Cactus	0.47	7.47
	BasketballDrive	0.59	6.90
	BQTerrace	0.28	6.39
C	BasketballDrill	0.46	7.29
	BQMall	0.53	6.65
	PartyScene	0.38	5.74
	RaceHorses	0.40	6.81
D	BasketballPass	0.60	7.82
	BQSquare	0.43	5.57
	BlowingBubbles	0.41	5.96
	RaceHorses	0.56	6.44
	Mean values	0.46	7.07

increased by less than 1% for the same objective quality. This small penalization can be justified by the limitations of the pre-analysis stage stated in Section 5, which result in sub-optimal decisions compared to the ones the encoder would have made. It should also be mentioned that these divergences are larger for smaller PUs such as 4×4, as they are typically harder to predict than the larger ones. Finally, the numeric results also show that there is some correlation between the BD-rate and the QP value used in the encoding, so that the higher the former, the higher the latter. This is probably related to the fact that the pre-analysis stage makes use of the original samples instead of the reconstructed ones.

As a consequence of the pre-analysis algorithm, the encoding time is redistributed as shown in Figure 4. These timing values have been extracted from the average profiles of all the tested sequences and QP values. As can be seen, a notable part of the time devoted to the intra module has been reduced, which is represented in the figure by the blue dotted bar. The reduced part corresponds to the first step of the 3-step intra algorithm implemented in HM, which is now performed in the pre-analysis algorithm. It can be seen that the decrement in time experienced by the encoder is larger than the increment introduced by the pre-analysis algorithm. It has to be noted that this small pre-analysis overhead does

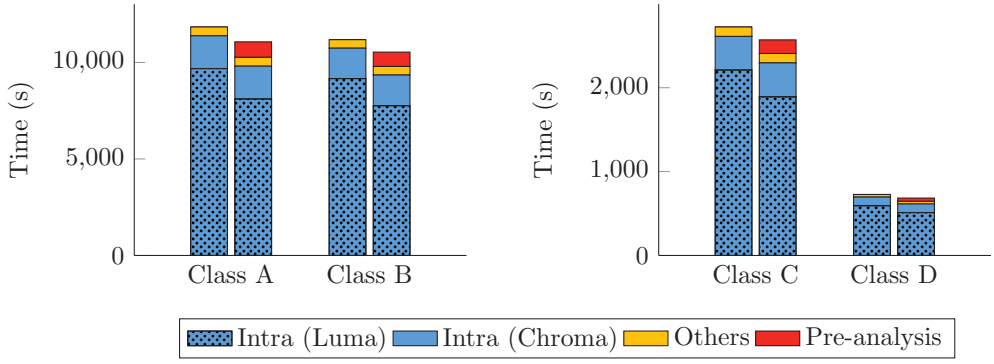


Figure 4: Profiling of the baseline (left side bar) compared to the proposal (right side bar)

not involve any latency, as the achieved time reduction is larger than the time spent on this stage.

7 Conclusions and Future Work

The advent of HEVC has opened the door to new applications and video formats such as UHD with the aim of fulfilling the demands for quality of experience of the market. However, the improved coding efficiency of HEVC has also led to a notable increment in the computational complexity of the encoder. As a consequence, HEVC-based encoders require efficient algorithms to achieve real-time encoding. Accordingly, this paper proposes a pre-analysis algorithm designed to estimate the intra prediction cost for every PU and intra mode in a fast way. This information is later used in the encoder to select the intra modes to be tested, and thus speed up the encoding. Additionally, this paper also details some considerations that have been taken into account in the design of the pre-analysis algorithm given the particularities of HEVC.

An experimental evaluation of the algorithm has shown that 7.07% of the encoding time can be saved on average with a negligible impact of 0.46% in BD-rate.

Future works include combining our prior work about pre-analysis algorithms to support both inter and intra modules in the encoder side. In this way, the encoder would take full advantage of this technique for other coding configurations such as Random Access. Furthermore, it could be possible to determine whether the encoder should only check inter or intra for a given CU with the help of the predicted costs. In a higher layer of abstraction, it could be also possible to determine the CTU partitioning based on these costs.

Acknowledgements

This work was jointly supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER funds) under the project TIN2015-66972-C5-2-R, and by the Spanish Ministry of Education, Culture and Sports under the grant FPU13/04601.

References

[1] ISO/IEC and ITU-T, *Advanced Video Coding for Generic Audiovisual Services. ITU-T Recommendation H.264 and ISO/IEC 14496-10 (version 10)*, February 2016.

[2] ISO/IEC and ITU-T, *High Efficiency Video Coding (HEVC). ITU-T Recommendation H.265 and ISO/IEC 23008-2 (version 3)*, April 2015.

[3] G. J. Sullivan, J.-R. Ohm, Woo-Jin Han, and T. Wiegand, *Overview of the High Efficiency Video Coding (HEVC) Standard*, IEEE Trans. Circuits Syst. Video Technol., 22 (2012), pp. 1649–1668.

[4] J.-R. Ohm, G. J. Sullivan, H. Schwarz, Thiow Keng Tan, and T. Wiegand, *Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC)*, IEEE Trans. Circuits Syst. Video Technol., 22 (2012), pp. 1669–1684.

[5] Chi Ching Chi, M. Álvarez-Mesa, B. Juurlink, G. Clare, F. Henry, S. Pateux, and T. Schierl, *Parallel Scalability and Efficiency of HEVC Parallelization Approaches*, IEEE Trans. Circuits Syst. Video Technol., 22 (2012), pp. 1827–1838.

[6] Chenggang Yan, Yongdong Zhang, Jizheng Xu, Feng Dai, Jun Zhang, Qionghai Dai, and Feng Wu, *Efficient Parallel Framework for HEVC Motion Estimation on Many-Core Processors*, IEEE Trans. Circuits Syst. Video Technol., 24 (2014), pp. 2077–2089.

[7] S. Radicke, J.-U. Hahn, Qi Wang, and C. Grecos, *Bi-predictive motion estimation for HEVC on a graphics processing unit (GPU)*, IEEE Trans. Consum. Electron., 60 (2014), pp. 728–736.

[8] Seunghyun Cho and Munchurl Kim, *Fast CU Splitting and Pruning for Suboptimal CU Partitioning in HEVC Intra Coding*, IEEE Trans. Circuits Syst. Video Technol., 23 (2013), pp. 1555–1564.

[9] Liquan Shen, Zhaoyang Zhang, and Ping An, *Fast CU size decision and mode decision algorithm for HEVC intra coding*, IEEE Trans. Consum. Electron., 59 (2013), pp. 207–213.

[10] Wie Jiang, Hanjie Ma, and Yaowu Chen, *Gradient based fast mode decision algorithm for intra prediction in HEVC*, in 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), April 2012, pp. 1836–1840.

[11] Hao Zhang and Zhan Ma, *Fast Intra Mode Decision for High Efficiency Video Coding (HEVC)*, IEEE Trans. Circuits Syst. Video Technol., 24 (2014), pp. 660–668.

[12] *x264 Source Code Repository*. `http://git.videolan.org/git/x264.git`.

[13] J. Garrett-Glaser, *A Novel Macroblock-Tree Algorithm for High-Performance Optimization of Dependent Video Coding in H.264/AVC*, tech. rep., Department of Computer Science. Harvey Mudd College, 2009.

[14] *x265 Source Code Repository*. `http://hg.videolan.org/x265`.

[15] Lin Sun, O. C. Au, Cong Zhao, and F. H. Huang, *Rate distortion modeling and adaptive rate control scheme for high efficiency video coding (HEVC)*, in IEEE International Symposium on Circuits and Systems (ISCAS), June 2014, pp. 1933–1936.

[16] T. Mallikarachchi, A. Fernando, and H. K. Arachchi, *Efficient coding unit size selection based on texture analysis for HEVC intra prediction*, in IEEE International Conference on Multimedia and Expo (ICME), July 2014, pp. 1–6.

[17] G. Cebrián-Márquez, Chi Ching Chi, J. L. Martínez, P. Cuenca, M. Álvarez-Mesa, S. Sanz-Rodríguez, and B. Juurlink, *Reducing HEVC Encoding Complexity Using Two-Stage Motion Estimation*, in IEEE International Conference on Visual Communications and Image Processing (VCIP), December 2015.

[18] G. Cebrián-Márquez, J. L. Martínez, and P. Cuenca, *A Pre-Analysis Algorithm for Fast Motion Estimation in HEVC*, in IEEE International Conference on Image Processing (ICIP), September 2016.

[19] *HEVC Test Model (HM) Reference Software*. `https://hevc.hhi.fraunhofer.de/`.

[20] F. Bossen, *Common Test Conditions and Software Reference Configurations*, Tech. Rep. JCTVC-L1100, January 2013.

[21] G. Bjøntegaard, *Calculation of average PSNR differences between RD-curves*, Tech. Rep. VCEG-M33, ITU-T Video Coding Experts Group (VCEG), 2001.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Exploiting Multi-level Parallelism on a Many-core System for the Application of Hyperheuristics to a Docking Problem

José M. Cecilia[1], José-Matías Cutillas-Lozano[2], Domingo Giménez[2] and Baldomero Imbernón[1]

[1] *Polytechnic School, Catholic University of San Antonio of Murcia (UCAM), Spain*

[2] *Departamento de Informática y Sistemas, University of Murcia, Spain*

emails: jmcecilia@ucam.edu, josematias.cutillas@um.es, domingo@um.es, bimbernon@alu.ucam.edu

Abstract

The solution of Protein-Ligand Docking Problems can be approached through metaheuristics, and satisfactory metaheuristics can be obtained with hyperheuristics searching in the space of metaheuristics implemented inside a parameterized schema. These hyperheuristics apply several metaheuristics, resulting in high computational costs. To reduce execution times, a shared-memory schema of hyperheuristics is used with four levels of parallelism, two for the hyperheuristic and two for the metaheuristics. The parallel schema is executed in a many-core system in "native mode", and the four-level parallelism allows us to take full advantage of the massive parallelism offered by this architecture and obtain satisfactory fitness and an important reduction of the execution time.

Key words: Parameterized metaheuristic schemas, Parallel metaheuristics, Hyperheuristics, Many-core system, Protein-Ligand Docking Problem

1 Introduction

In the Protein-Ligand Docking Problem (PLDP) a scoring function is optimized to obtain the position at which a ligand best matches a given protein. Different scoring functions can be used [8]. In this work the scoring function is computed through the Lennard-Jones potential, obtained as the sum of the interactions of each atom, i, of an active site of the protein with each atom, j, of the ligand:

$$V(i,j) = 4\epsilon \left(\left(\frac{\sigma}{r(i,j)} \right)^{12} - \left(\frac{\sigma}{r(i,j)} \right)^{6} \right) \quad (1)$$

where σ and ϵ are empirical constants of the model, and $r(i,j)$ is the distance between atoms i and j.

The PLDP can be seen as a problem of searching for the values of the degrees of freedom (six) that globally minimize the scoring function. The values of the movements and rotations of the ligand can be approached with metaheuristics, and hyperheuristics can be used for the selection of satisfactory metaheuristics for the PLDP.

Hyperheuristics select satisfactory metaheuristics or build new ones by combining basic metaheuristics for a particular problem [3, 6]. In our approach, we use a parameterized schema of metaheuristics [2], which facilitates the selection of metaheuristics or their combinations by selecting numerical values for metaheuristic parameters in the schema. It considers a set of basic functions whose instantiation determines the particular metaheuristic being implemented. Hyperheuristics can be developed on top of the parameterized schema [4], and they search automatically in the space of metaheuristics for a satisfactory metaheuristic for a particular problem.

When hyperheuristics are used to select satisfactory metaheuristics the execution time increases significantly, so high performance computing strategies are compulsory.

This paper analyses the use of massive parallelism in a MIC (Xeon Phi) for exploiting multilevel parallelism in hyperheuristics working on top of parameterized metaheuristics when applied to the PLDP. The multilevel approach allows the application of parallelism at hyperheuristic and metaheuristic levels, and the optimal combination of the number of threads to use at each level should be selected.

The rest of the paper is organized as follows. Section 2 gives the basis of the shared-memory, parameterized metaheuristic schema which is used for the development of hyperheuristics in a MIC. Section 3 discuses the experimental results obtained with the application of hyperheuristics to an instance of the PLDP. Section 4 concludes and outlines possible research directions.

2 A parallel, parameterized metaheuristic schema

Our approach for parallelizing metaheuristics consists of the parallelization of a unified parameterized metaheuristic schema (Algorithm 1). The hyperheuristics and the metaheuristics they search for are implemented with the same schema, and so they are parallelized in the same way. In the schema, $ParamX$ represents the metaheuristic parameters and $ThreadsX$ the parallelism parameters for each basic function.

Algorithm 1 Parameterized shared-memory metaheuristic schema

```
Initialize(S,ParamIni,ThreadsIni)
while (not EndCondition(S,ParamEnd)) do
  SS=Select(S,ParamSel)
  SS1=Combine(SS,ParamCom,ThreadsCom)
  SS2=Improve(SS1,ParamImp,ThreadsImp)
  S=Include(SS2,ParamInc,ThreadsInc)
end while
```

The basic functions in the schema can be implemented in different ways, and the number of parameters and their meanings could change. We are not interested here in an in-depth discussion of the metaheuristic and the parallelism parameters. Interested readers can consult [1] (parallelism parameters) and [2] (metaheuristic parameters). Some of the routines are parallelized with only one level of parallelism, and in other functions two levels are used.

The parallelization based on the shared-memory paradigm [1] is adapted here to a many-core system for the massive parallelization of the schema, which allows four levels of parallelism, including parallelism in the hyperheuristic and in the metaheuristics being selected. The schema is executed directly on the coprocessor without offloading from a host system, which is known as running in "native mode".

3 Experimental results

The PLDP can be seen as the problem of searching for the values of the degrees of freedom (six) that globally minimize the scoring function. The values of the movements and rotations of the ligand can be approached with metaheuristics. The application of parallelism to this problem is analyzed in [5], and the application to the problem of hyperheurstics in a MIC is analyzed here.

Experiments were carried out in a Many Integrated Core (MIC) Intel Xeon Phi with 57 cores at 1.1 GHz based on Pentium (x86), with each core supporting 4 hardware threads, with a bidirectional ring bus and up to 6 GBytes GDDR5.

The protein PDB:2BSM, from the well-known Protein Data Bank [7], is used for the experiments. Due to the high size of the problem, hyperheuristics are applied here to this reduced instance. The study shows the influence on the execution times of the division of threads between the four levels of parallelism in the MIC architecture, in order to enhance the performance of our hyperheuristic schema when massive parallelism is required. Due to the high computational cost of applying hyperheuristics to the PLDP and because the study was not focused on optimizing fitness, a Reduced Hyperheuristic (Hre) was used in the experiments. Table 1 shows the values of the metaheuristic parameters considered. As

mentioned, the number and meaning of the parameters depend on the particular implementation of the parameterized schema. Our implementation considers eighteen metaheuristic parameters (five in the initialization, two in the selection, three for combination, six for the improvement, and two for the inclusion). Some elements are initially generated (INEIni), and some are selected for the successive steps (FNEIni). A certain percentage of the generated elements (PEIIni) is improved with a given intensity (IIEIni) and with a short tabu memory (STMIni); and the corresponding parameters are used for improvement of elements after combination (PEIImp, IIEImp and SMIImp) and diversification (PEDImp, IDEImp and SMDImp). In each iteration of the algorithm, some elements are selected from the best (NBESel) and the worst (NWWSel) ones, and combinations between pairs of best, worst and best-worst elements (NBBCom, NWWCom and NBWCom) are made. Some of the best elements are selected for the following iteration (NBEInc), with the use of a long-term tabu memory (LTMInc). Due to the high computational cost of the hyperheuristic, low values were fixed for most of the hyperheuristic parameters, and more threads are devoted to the computations with the metaheuristics the hypeheuristic experiments with.

Table 1: Values of the metaheuristic parameters for the Reduced Hyperheuristic (Hre) used in the experiments.

INEIni	FNEIni	PEIIni	IIEIni	STMIni	NBESel
5	5	50	3	0	3
NWESel	NBBCom	NBWCom	NWWCom	PEIImp	IIEImp
2	2	3	2	50	3
SMIImp	PEDImp	IDEImp	SMDImp	NBEInc	LTMInc
0	10	5	0	3	5

Different values of the parallelism parameters for the Reduced Hyperheuristic in Table 1 are considered for the executions in the range between 20 and 250 threads (ThT). The threads are spread over the hyperheuristic (ThH) and the metaheuristics (ThM). The product of the number of threads in the hyperheuristic and the metaheuristics is equal to the total ($\text{ThM} \cdot \text{ThH} = \text{ThT}$). The ThH combinations for ThT=20 are shown in Table 2. For example, for the 5-th series (ThH_20_5), with a total of ThT=20 threads, when the initial generation of the reference set is executed for the hyperheuristic with ThH=5 threads, the metaheuristics are executed with ThM=20/5=4 threads. The values of the metaheuristic parameters for the metaheuristics are higher than for the hyperheuristic, and this results in coarse grained parallelism, for which a large number of threads is preferred in the first level of the two level routines, and the number of threads in the second level of the metaheuristics is fixed to 1.

Table 3 shows the experimental times (in seconds) obtained when applying the Reduced

Table 2: Number of threads of one and two levels of parallelism for the execution of the Reduced Hyperheuristic in Table 1, with the total number of threads set to 20. *ThH_ThT_series* represents *threads used in the hyperheuristic_total number of threads_series of the experiment*, and the remaining threads are assigned to the metaheuristics is at low level (ThM), with ThM · ThH = ThT.

		One-level parallel routines			Two-level parallel routines				
ThH_ThT_series		TGEIni	TCPCom	TIEInc	TI_Ini	TR_Imp	TC_Imp	TR_Div	TC_Div
ThH_20_1	p_1	1	1	1	1	1	1	1	1
	p_2	-	-	-	2	2	2	1	1
ThH_20_2	p_1	2	2	2	1	1	1	2	2
	p_2	-	-	-	2	2	2	1	1
ThH_20_3	p_1	5	5	5	2	2	2	5	5
	p_2	-	-	-	2	2	2	1	1
ThH_20_4	p_1	5	5	5	1	1	1	1	1
	p_2	-	-	-	5	5	5	2	2
ThH_20_5	p_1	5	5	5	1	1	1	2	2
	p_2	-	-	-	5	5	5	2	2
ThH_20_6	p_1	10	10	10	2	2	2	5	5
	p_2	-	-	-	5	5	5	2	2
ThH_20_7	p_1	10	10	10	1	1	1	1	1
	p_2	-	-	-	10	10	10	5	5
ThH_20_8	p_1	10	10	10	1	1	1	2	2
	p_2	-	-	-	10	10	10	5	5
ThH_20_9	p_1	20	20	20	2	2	2	4	4
	p_2	-	-	-	10	10	10	5	5

Hyperheuristic in Table 1 to search for satisfactory metaheuristics for the problem PLDP in Xeon Phi. Due to the small parameter values managed by the Hre (NFEIni=5 among others), values of ThT greater than 20 have the same thread combinations as ThH_20 for all the series in Table 2, so the influence of the threads of the metaheuristic is given for a fixed configuration of the Hre threads (ThH_20 to ThH_250 series). The lowest times are achieved in most cases for the series comprised between ThH_ThT_3 and ThH_ThT_5, which is not surprising because the number of threads of the hyperheuristic is similar to the size of the population parameters, with INEIni = FNEIni in that series.

Table 3: Execution time in seconds for the Hre in Table 1 and various thread combinations in Table 2 applied to the PLDP in Xeon Phi. The sequential time in Xeon Phi was 7983 seconds.

	total number of threads (ThT)					
thread combination	20	50	100	150	200	250
ThH_ThT_1	795	428	346	367	366	361
ThH_ThT_2	782	508	388	296	327	422
ThH_ThT_3	940	495	354	283	282	338
ThH_ThT_4	765	519	359	339	324	393
ThH_ThT_5	950	469	343	375	390	397
ThH_ThT_6	1021	530	396	401	345	365
ThH_ThT_7	1449	696	398	331	477	357
ThH_ThT_8	1436	734	407	407	308	442
ThH_ThT_9	1512	910	459	416	318	**265**

The advantage of using parallelism is clear here, with a maximum speed-up of approximately 30 with respect to the sequential execution in Xeon Phi and with the total number of threads close to the maximum available. So, for this computationally demanding problem, the massive parallelism of MIC is well exploited with the parallel, parameterized metaheuristic schema.

4 Conclusions and future work

A shared-memory schema of hyperheuristics is used to select satisfactory metaheuristics to be applied to a molecular docking problem. Due to the high computational cost of the hyperrheuristic, the parallel schema was executed in a many-core system in "native mode" with four levels of parallelism, which allows us to take full advantage of the massive parallelism offered by this architecture, obtaining an important reduction in execution times.

The best results are obtained with a relatively low number of threads assigned to the hyperheuristic, and the rest are allocated to the low level metaheuristic, with the total number of threads close to the maximum available.

For higher reductions of the execution time it would be necessary to combine parallelism in the multicore host and the MIC coprocessor.

As future research lines, the parameterized schema could be applied to other optimization problems. The inclusion of new basic metaheuristics, for example, Ant Colony Optimization or Particle Swarm Optimization, is also contemplated. Similar parameterized, parallel metaheuristic schemas should be developed for GPU and in heterogeneous clusters comprising nodes of multicores + multiple GPU or MIC. The use of large, heterogeneous clusters would be of special interest for the application of hyperheuristics with large reference sets or with a high fitness function cost, as in the case of the molecular docking problems.

Acknowledgements

This work was supported by the Spanish MINECO, as well as European Commission FEDER funds, under grant TIN2015-66972-C5-3-R.

References

[1] F. Almeida, D. Giménez, and J.-J. López-Espín. A parameterized shared-memory schema for parameterized metaheuristics. *The Journal of Supercomputing*, 58(3):292–301, 2011.

[2] F. Almeida, D. Giménez, J.-J. López-Espín, and M. Pérez-Pérez. Parameterised schemas of metaheuristics: basic ideas and applications with Genetic Algorithms, Scatter Search and GRASP. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 43(3):570–586, 2013.

[3] E. K. Burke, M. Hyde, G. Kendall, G. Ochoa, E. Ozcan, and J. R. Woodward. *A Classification of Hyper-heuristic Approaches*, volume 146, pages 449–468. Springer, 2010. In Michel Gendreau, Jean-Yves Potvin, editors, Handbook of Metaheuristics.

[4] J.-M. Cutillas-Lozano, D. Giménez, and F. Almeida. Hyperheuristics based on parameterized metaheuristic schemas. In *Genetic and Evolutionary Computation Conference (GECCO'15)*, pages 361–368, 2015.

[5] B. Imbernón, J. M. Cecilia, and D. Giménez. Enhancing Metaheuristic-based Virtual Screening Methods on Massively Parallel and Heterogeneous Systems. In *Proceedings of the 7th International Workshop on Programming Models and Applications for Multicores and Manycores, PMAM@PPoPP*, pages 50–58, 2016.

[6] E. Ozcan, B. Bilgin, and E. Korkmaz. A comprehensive analysis of hyper-heuristics. *Intelligent Data Analysis*, 12(1):3–23, 2008.

[7] Protein Data Bank. *Nature New Biol*, 233:223, 1971.

[8] E. Yuriev, M. Agostino, and P. A. Ramsland. Challenges and Advances in Computational Docking: 2009 in Review. *Journal of Molecular Recognition*, 24(2):149–164, 2011.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A non-singular and positive Bhattacharya method for the numerical modeling of the dewetting process of thin films

Axel Chávez-Guzmán[1], I. E. Medina-Ramírez[2] and J. E. Macías-Díaz[3]

[1] *Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes*

[2] *Departamento de Química, Universidad Autónoma de Aguascalientes*

[3] *Departamento de Matemáticas y Física, Universidad Autónoma de Aguascalientes*

emails: , iemedina@correo.uaa.mx, jemacias@correo.uaa.mx

Abstract

In this work, we provide a discrete mathematical system to model the dynamics of the thickness of two-dimensional thin films subject to a dewetting process. The model under considerations is a degenerate generalization of the classical thin-film equation, and considers the inclusion of a singular potential. The analytical model is discretized using a modification of the exponential method employed by Bhattacharya and co-workers. Our correction yields an explicit numerical technique that is non-singular with respect to zero solutions of the mathematical model, and that is capable of preserving the non-negative character of the approximations. In addition, the explicit nature of our approach results in an economic computer implementation which produces fast simulations. *Key words:*

Degenerate thin-film equation, modified Bhattacharya exponential method, preservation of non-negativity, computationally efficient numerical technique

1 Introduction

Throughout this work, we let A, ϵ, η, κ and σ be positive numbers. Let Ω represent an open, bounded and connected domain of $\mathbb{R}^2$ with boundary $\partial\Omega$, and let $\overline{\Omega}$ represent its closure in the standard topology. Assume that $h : \overline{\Omega} \times \overline{\mathbb{R}^+} \to \mathbb{R}$ is a function with continuous derivatives up to the fourth order that satisfies the initial-boundary-value problem

$$\begin{aligned} &\eta\frac{\partial h}{\partial t}(\mathbf{x},t) - \nabla\cdot(m(h(\mathbf{x},t))\nabla p(\mathbf{x},t)) = 0, \\ &\text{subject to } \begin{cases} h(\mathbf{x},t) = \varphi(\mathbf{x}), & \forall(\mathbf{x},t)\in\Omega\times\{0\}, \\ \hat{\mathbf{n}}\cdot\nabla h(\mathbf{x},t) = 0, & \forall(\mathbf{x},t)\in\partial\Omega\times\mathbb{R}^+, \end{cases} \end{aligned} \tag{1}$$

 ISBN: 978-84-608-6082-2

for every $(\mathbf{x},t)\in\Omega\times\mathbb{R}^+$. In our context, Ω will represent a spatial domain and the variable t will denote time. Meanwhile, the operators $\nabla\cdot$, ∇ and ∇^2 will be used to represent the two-dimensional divergence, the gradient and the Laplacian operators in the spatial variables, respectively. We will assume that $\varphi:\Omega\to\mathbb{R}$ will be a nonnegative function, and p will be given by

$$p(\mathbf{x},t)=-\sigma\nabla^2 h(\mathbf{x},t)+V'(h(\mathbf{x},t)), \tag{2}$$

for each $\mathbf{x}\in\Omega$ and $t>0$. Finally,

$$m(h(\mathbf{x},t)) = \frac{h(\mathbf{x},t)^3}{3}, \tag{3}$$

$$V(h(\mathbf{x},t)) = \frac{\epsilon}{h^8(\mathbf{x},t)}-\frac{A}{12\pi h^2(\mathbf{x},t)}. \tag{4}$$

From a physical perspective, the partial differential equation in (1) describes the spatial and temporal evolution of the rupture of ultrathin films [1]. The function h provides the thickness of the thin film at each point $\mathbf{x}$ of a two-dimensional substrate Ω, and each time $t\geq 0$. In that context, the constant σ represents the surface tension, η is the constant viscosity parameter, A is the Hamaker constant of polystyrene on a compound, and ϵ is the strength of the potential V. Meanwhile, p is the augmented Laplace pressure, and m is the specific mobility coefficient [2]. Mathematically, the model (1) is a generalized thin-film equation for which globally non-negative solutions exist [3, 4, 5]. This fact is physically relevant in view that h represents the (non-negative) thickness of a thin film. Assuming the h is a non-negative solution of (1), we can divide both sides of this equation by $h(\mathbf{x},t)+\kappa$, whence we readily obtain the equivalent partial differential equation

$$\eta\frac{\partial}{\partial t}\ln h(\mathbf{x},t)=\frac{\nabla\cdot(m(h(\mathbf{x},t))\nabla p(\mathbf{x},t))}{h(\mathbf{x},t)+\kappa}, \tag{5}$$

for each $\mathbf{x}\in\Omega$ and each $t>0$.

Analytically, the resolution of a problem governed by the partial differential equation of (1) is as complicated as solving a problem ruled by (5). However, the numerical discretization of the latter partial differential equation results in a new family of techniques that possesses efficiency properties and preserves some of the features of the relevant solutions. Indeed, due to the difficulty of determine solutions of the thin-film equation (1), we would ideally require for a method to fulfill the following computational characteristics: to be computationally fast, to be easy to implement in any computer program, to be able to handle fine grid meshes, and to require a reasonable amount of computer memory. In addition to these properties, we also want for the method to possess the next numerical characteristics: to preserve the non-negativity of solutions, to be robust with respect to zero solutions, to be stable, and to be convergent. In the present work, we design a numerical method which satisfies most of the computational and numerical characteristics mentioned above.

2 Finite-difference scheme

For each positive integer $P \geq 4$, define the finite sets $I_P = \{1, \ldots, P-1\}$, $\overline{I}_P = I_P \cup \{0, P\}$ and $\operatorname{int} I_P = I_P \setminus \{1, P-1\}$. Let K, M and N be positive integers. We will restrict our attention to spatial domains of the form $\Omega = [a,b] \times [c,d]$ of $\mathbb{R}^2$, where $a < b$ and $c < d$. We will fix regular partitions $(x_i)_{i=0}^M$ and $(y_j)_{j=0}^N$ of the spatial intervals $[a,b]$ and $[c,d]$, respectively, with partition norms denoted by Δx and Δy, respectively. Let T be a positive number, and let $(t_k)_{k=0}^K$ a (not necessarily uniform) partition of the temporal interval $[0,T]$, and let $\Delta t_k = t_{k+1} - t_k$ for each $k = I_K \cup \{0\}$.

Let $D = \{(x_i, y_j, t_k) : i \in \overline{I}_M, j \in \overline{I}_N, k \in \overline{I}_K\}$, and let $u : D \to \mathbb{R}$ be a function. We define the following forward-difference operators for each $i \in I_M$, $j \in I_N$ and $k \in I_K$:

$$\delta_x^{\pm} u(x_i, y_j, t_k) = \frac{u(x_{i\pm 1}, y_j, t_k) - u(x_i, y_j, t_k)}{\pm \Delta x}, \tag{6}$$

$$\delta_y^{\pm} u(x_i, y_j, t_k) = \frac{u(x_i, y_{j\pm 1}, t_k) - u(x_i, y_j, t_k)}{\pm \Delta y}, \tag{7}$$

$$\delta_t u(x_i, y_j, t_k) = \frac{u(x_i, y_j, t_{k+1}) - u(x_i, y_j, t_k)}{\Delta t_k}, \tag{8}$$

$$\mu_x^{\pm} u(x_i, y_j, t_k) = \frac{u(x_{i\pm 1}, y_j, t_k) + u(x_i, y_j, t_k)}{2}, \tag{9}$$

$$\mu_y^{\pm} u(x_i, y_j, t_k) = \frac{u(x_i, y_{j\pm 1}, t_k) + u(x_i, y_j, t_k)}{2}, \tag{10}$$

The first three operators provide first-order approximations to the partial derivative of u with respect to x, y and y, respectively, at the point (x_i, y_j, t_k). We also define the following quadratic approximation of the second-order partial derivatives of u with respect to z, for $z = x, y$:

$$\delta_{zz} u(x_i, y_j, t_k) = \frac{\delta_z^+ u(x_i, y_j, t_k) - \delta_z^- u(x_i, y_j, t_k)}{\Delta z}. \tag{11}$$

Clearly, the operator $\widehat{\nabla}^2 u(x_i, y_j, t_k) = [\delta_{xx} + \delta_{yy}] u(x_i, y_j, t_k)$ provides a second-order approximation for the Laplacian of u at the point $u(x_i, y_j, t_k)$.

We will employ this notation to define approximations of the differential operators introduced in Section 1. For instance, we will approximate the operator p using the discrete operator

$$\widehat{p}(x_i, y_j, t_k) = -\sigma \widehat{\nabla}^2 h(x_i, y_j, t_k) + V'(h(x_i, y_j, t_k)), \tag{12}$$

for each $(i, j, k) \in I_M \times I_N \times I_K$. It is worth noticing here that the term $V'(h(\mathbf{x}, t))$ will be approximated in an exact form in order to reduce the computer time of the simulations. On the other hand, let

$$d(\mathbf{x}, t) = \nabla \cdot (m(h(\mathbf{x}, t)) \nabla p(\mathbf{x}, t). \tag{13}$$

The finite-difference approximation for the nonlinear diffusion operator at the point (x_i, y_j, t_k) will be provided by the expression

$$\widehat{d}(x_i, y_j, t_k) = \sum_{z=x,y} \frac{\widehat{d}_z^+(x_i, y_j, t_k) - \widehat{d}_z^-(x_i, y_j, t_k)}{\Delta z}, \tag{14}$$

for each $i \in \operatorname{int} I_M$, $j \in \operatorname{int} I_N$ and $k \in I_K \cup \{0\}$. Here, for each $z = x, y$,

$$\widehat{d}_z^{\pm}(x_i, y_j, t_k) = m(\mu_z^{\pm} h(x_i, y_j, t_k))\delta_z^{\pm}\widehat{p}(x_i, y_j, t_k). \tag{15}$$

Using this notation, the finite-difference discretization used to approximate the solutions of (5) is given by the following system of equations

$$\eta\delta_t \ln(u(x_i, y_j, t_k) + \kappa) = \frac{\widehat{d}(x_i, y_j, t_k)}{h(x_i, y_j, t_k) + \kappa}, \tag{16}$$

for each $i \in \operatorname{int} I_M$, $j \in \operatorname{int} I_N$ and $k \in I_K \cup \{0\}$. Of course, initial and boundary conditions must be imposed in order for the method to be completely defined.

References

[1] Jürgen Becker, Günther Grün, Ralf Seemann, Hubert Mantz, Karin Jacobs, Klaus R Mecke, and Ralf Blossey. Complex dewetting scenarios captured by thin-film models. *Nature Materials*, 2(1):59–63, 2003.

[2] Stephen H Davis et al. On the motion of a fluid-fluid interface along a solid surface. *Journal of Fluid Mechanics*, 65(01):71–95, 1974.

[3] Francisco Bernis and Avner Friedman. Higher order nonlinear degenerate parabolic equations. *Journal of Differential Equations*, 83(1):179–206, 1990.

[4] AL Bertozzi and M Pugh. The lubrication approximation for thin viscous films: regularity and long-time behavior of weak solutions. *Communications on Pure and Applied Mathematics*, 49(2):85–123, 1996.

[5] Roberta Dal Passo, Harald Garcke, and Günther Grün. On a fourth-order degenerate parabolic equation: global entropy estimates, existence, and qualitative behavior of solutions. *SIAM Journal on Mathematical Analysis*, 29(2):321–342, 1998.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Rovibrational States of Wigner Molecules in Spherically Symmetric Confining Potentials

Jerzy Cioslowski[1]

[1] *Institute of Physics, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland*

email: jerzy@wmf.univ.szczecin.pl

Abstract

The strong-localization limit of three-dimensional Wigner molecules, in which repulsively interacting particles are confined by a vanishingly weak spherically symmetric potential, is investigated. An explicit prescription for computation of rovibrational wavefunctions and energies that are asymptotically exact at this limit is presented. The performance of the new formalism is illustrated with the three- and four-electron harmonium atoms at their strong-correlation limits.

Key words: Wigner molecules, strong-correlation, harmonium atoms

In systems with external potentials that are sufficiently weak to allow for large interelectron distances yet steep enough to prevent ionization, localization of electrons away from the minima of the external potential gives rise to species known as Wigner molecules [1, 2]. At the limit of an infinitesimally weak confinement, the electronic states of these species are *de facto* properly antisymmetrized parity- and spin-adapted eigenstates of a rovibrational Hamiltonian, derivation of which shares some of its steps with the formalism of Louck [3] that avoids references to classical quantities inherent in more conventional approaches [4]. However, the treatment of species composed exclusively of electrons differs in several aspects from that pertaining to their counterparts comprising nuclei. First of all, since only the leading asymptotics of the wavefunctions of individual rovibrational states and their energy ordering are of interest here, both the Watson term [5] and the contributions due to the vibrational angular momenta and the dependence of the inertia tensor on the vibrational coordinates can be safely neglected. Second, separation of the translational degrees of freedom has to be abandoned due to the center-of-mass motions (the Kohn modes [6])

 ISBN: 978-84-608-6082-2

being harmonic rather than free-particle-like. Finally, some trivial simplifications ensue due to the equal masses of all the particles in question.

Consider a Wigner molecule comprising N electrons confined by a spherically symmetric external potential. The rotational invariance of the potential energy $V(\vec{R})$ (which is the sum of the confinement and electron-electron repulsion energies) implies $V(\vec{R}) = V(\boldsymbol{U}\vec{R})$, where $\vec{R} \equiv (\vec{r}_1, \ldots, \vec{r}_N)$ is a supervector of N position vectors $\{\vec{r}_j\}$ and $\boldsymbol{U}\vec{R} \equiv (\boldsymbol{u}\vec{r}_1, \ldots, \boldsymbol{u}\vec{r}_N)$, for any unitary matrix $\boldsymbol{u}$. $V(\vec{R})$ has the global minimum at $\vec{\mathfrak{R}}^0 \equiv (\vec{\mathfrak{r}}^0_1, \ldots, \vec{\mathfrak{r}}^0_N)$ that is oriented in such a way that the corresponding inertia tensor $\boldsymbol{I} = \sum_{j=1}^{N}[\,(\vec{\mathfrak{r}}^0_j \cdot \vec{\mathfrak{r}}^0_j)\mathbf{1} - \vec{\mathfrak{r}}^0_j \otimes \vec{\mathfrak{r}}^0_j\,]$ (where $\mathbf{1}$ is the unit matrix) is diagonal (i.e. $[\boldsymbol{I}]_{kk'} = \delta_{kk'}\, I_k$; here and in the following, $[\boldsymbol{a}]_k$ and $[\boldsymbol{A}]_{kk'}$ denote, respectively, the kth and kk'th element of the vector $\boldsymbol{a}$ and the matrix $\boldsymbol{A}$). Consequently,

$$V(\vec{\mathfrak{R}}) \approx V(\vec{\mathfrak{R}}^0) + \frac{1}{2}\sum_{p=1}^{3N-3} \varepsilon_p \left[\, \vec{\Xi}_p \cdot (\vec{\mathfrak{R}} - \vec{\mathfrak{R}}^0) \,\right]^2 \quad , \tag{1}$$

where $\{\varepsilon_p\}$ and $\{\vec{\Xi}_p\} \equiv \{(\vec{\xi}_{p1}, \ldots, \vec{\xi}_{pN})\}$ are, respectively, the eigenvalues and the normalized eigenvectors of the Hessian of $V(\vec{\mathfrak{R}})$ at $\vec{\mathfrak{R}} = \vec{\mathfrak{R}}^0$ (here and in the following, the symbol $\approx$ implies smallness of the norm $\|\vec{\mathfrak{R}} - \vec{\mathfrak{R}}^0\|$). The three vanishing eigenvalues, which do not enter the sum in the expansion (1), pertain to the eigenvectors $\{\vec{\Xi}_{3N-2}, \vec{\Xi}_{3N-1}, \vec{\Xi}_{3N}\}$ that describe infinitesimal rotations. Those are conveniently chosen as

$$\vec{\Xi}_{3N-3+k} = I_k^{-1/2}\,(\vec{e}_k \times \vec{\mathfrak{r}}^0_1, \ldots, \vec{e}_k \times \vec{\mathfrak{r}}^0_N) \quad , \tag{2}$$

where $\vec{e}_1 = (1,0,0)$, $\vec{e}_2 = (0,1,0)$, and $\vec{e}_3 = (0,0,1)$ are three unit vectors.

The above observations prompt introduction of the parameterization

$$\vec{r}_j = \boldsymbol{T}(\theta_1, \theta_2, \theta_3)\left(\vec{\mathfrak{r}}^0_j + \sum_{p=1}^{3N-3} q_p\, \vec{\xi}_{pj}\right) \tag{3}$$

for the position vectors $\{\vec{r}_j\}$ in terms of $3N-3$ normal displacements $\{q_p\}$ and the Euler angles $\{\theta_1, \theta_2, \theta_3\}$ of the rotation matrix (in the zyz/active convention)

$$\boldsymbol{T}(\theta_1, \theta_2, \theta_3) = \begin{pmatrix} \cos\theta_1\cos\theta_2\cos\theta_3 - \sin\theta_1\sin\theta_3 & -\cos\theta_1\cos\theta_2\sin\theta_3 - \sin\theta_1\cos\theta_3 & \cos\theta_1\sin\theta_2 \\ \sin\theta_1\cos\theta_2\cos\theta_3 + \cos\theta_1\sin\theta_3 & -\sin\theta_1\cos\theta_2\sin\theta_3 + \cos\theta_1\cos\theta_3 & \sin\theta_1\sin\theta_2 \\ -\sin\theta_2\cos\theta_3 & \sin\theta_2\sin\theta_3 & \cos\theta_2 \end{pmatrix} \tag{4}$$

that satisfy the constraints $\theta_1 \in [0, 2\pi]$, $\theta_2 \in [0, \pi]$, and $\theta_3 \in [0, 2\pi]$. Application of this parameterization leads to the familiar rovibrational Hamiltonian

$$\hat{H}_{rv} = V(\vec{\mathfrak{R}}^0)\,\hat{1} + \frac{1}{2}\sum_{p=1}^{3N-3}\left(-\frac{\partial^2}{\partial q_p^2} + \varepsilon_p\,\hat{q}_p^2\right) + \frac{1}{2}\sum_{k=1}^{3} I_k^{-1}\,\hat{P}_k^2 \quad , \tag{5}$$

 ISBN: 978-84-608-6082-2

where $\hat{P}_k$ is the kth component of the body-fixed angular momentum operator. The eigenenergies of $\hat{H}_{rv}$ read

$$E_{Lmn;\{\nu_p\}} = V(\vec{\mathfrak{R}}^0) + \sum_{p=1}^{3N-3} \varepsilon_p^{1/2} \left(\nu_p + \frac{1}{2}\right) + \epsilon_{Ln} \quad , \tag{6}$$

where the integer quantum numbers L, m, and n satisfy the conditions $L \geq 0$, $|m| \leq L$, and $|n| \leq L$, and the elements of the set $\{\nu_p\}$ of the vibrational quantum numbers are nonnegative integers. Neither the nth eigenvalue ϵ_{Ln} nor the corresponding normalized eigenvector $\boldsymbol{c}_{Ln}$ of the matrix $\boldsymbol{B}_L$ with the elements [7]

$$\begin{aligned}[\boldsymbol{B}_L]_{\mu\mu'} = & \left[\frac{I_1^{-1} + I_2^{-1}}{4}\, L(L+1) + \frac{2\, I_3^{-1} - (I_1^{-1} + I_2^{-1})}{4}\, \mu^2\right] \delta_{\mu,\mu'} + \frac{I_1^{-1} - I_2^{-1}}{8} \\ & \times \Big([L(L+1) - (\mu-1)(\mu-2)]^{1/2}\, [L(L+1) - \mu(\mu-1)]^{1/2}\, \delta_{\mu,\mu'+2} \\ & \quad [L(L+1) - (\mu+1)(\mu+2)]^{1/2}\, [L(L+1) - \mu(\mu+1)]^{1/2}\, \delta_{\mu,\mu'-2}\Big)\end{aligned} \tag{7}$$

(where $|\mu| \leq L$ and $|\mu'| \leq L$) depends on m hence, as expected, the eigenenergies of $\hat{H}_{rv}$ are m-independent. In the case of $I_1 = I_2$, $\boldsymbol{B}_L$ is diagonal and thus $[\boldsymbol{c}_{Ln}]_\mu = \delta_{n\mu}$. The eigenfunctions of $\hat{H}_{rv}$ are given by

$$\begin{aligned}\Psi_{Lmn;\{\nu_p\}}(\theta_1, \theta_2, \theta_3, \{q_p\}) = & \pi^{-3(N-1)/4}\, (I_1 I_2 I_3)^{-1/4} \left[\sum_{\mu=-L}^{L} [\boldsymbol{c}_{Ln}]_\mu\, D_{Lm\mu}(\theta_1, \theta_2, \theta_3)\right] \\ & \times \prod_{p=1}^{3N-3} \left[2^{-\nu_p/2}\, (\nu_p!)^{-1/2}\, \varepsilon_p^{1/8}\, H_{\nu_p}(\varepsilon_p^{1/4}\, q_p)\, \exp\left(-\frac{\varepsilon_p^{1/2}\, q_p^2}{2}\right)\right] \quad ,\end{aligned} \tag{8}$$

where

$$\begin{aligned}D_{Lm\mu}(\theta_1, \theta_2, \theta_3) = & \left[\frac{2L+1}{8\,\pi^2}\, \frac{(L+m)!\,(L-m)!}{(L+\mu)!\,(L-\mu)!}\right]^{1/2} \\ & \times P_{L-m}^{m-\mu, m+\mu}(\cos\theta_2)\, \left(\cos\frac{\theta_2}{2}\right)^{m+\mu} \left(\sin\frac{\theta_2}{2}\right)^{m-\mu} \exp[i\,(m\,\theta_1 + \mu\,\theta_3)]\end{aligned} \tag{9}$$

[in Eqs. (8) and (9), $H_n(t)$ and $P_n^{(\alpha,\beta)}(t)$ are, respectively, the Hermite and Jacobi polynomials].

The actual rovibrational states ensue from these primitive wavefunctions upon parity adaptation, incorporation of spin variables, and antisymmetrization, which is usually more facile for $\Psi_{Lmn;\{\nu_p\}}(\vec{R})$, i.e. for $\Psi_{Lmn;\{\nu_p\}}(\theta_1, \theta_2, \theta_3, \{q_p\})$ expressed in terms of the position

vectors. Computation of such wavefunctions requires inversion of the parameterization (3), which produces

$$q_p = \sum_{j=1}^{N} \left[\boldsymbol{T}(\theta_1, \theta_2, \theta_3)\, \vec{\xi}_{pj} \right] \cdot \vec{r}_j - \sum_{j=1}^{N} \vec{\xi}_{pj} \cdot \vec{\mathfrak{r}}_j^{\,0} \tag{10}$$

thanks to the orthonormality of $\{\vec{\Xi}_p\}$ for $1 \le p \le 3N-3$, whereas the rotation matrix obtains from the expression [8]

$$\boldsymbol{T}(\theta_1, \theta_2, \theta_3) = \frac{2\,\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega}\,\boldsymbol{\Omega}^{\boldsymbol{T}} - [\kappa^2 + \mathrm{Tr}(\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega})]\,\boldsymbol{\Omega}^{\boldsymbol{T}} - 2\,\kappa\,(\det\boldsymbol{\Omega})\,\boldsymbol{\Omega}^{-\boldsymbol{1}}}{2\,\det\boldsymbol{\Omega} - \kappa\,[\kappa^2 - \mathrm{Tr}(\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega})]} \quad , \tag{11}$$

where $\boldsymbol{\Omega} \equiv \boldsymbol{\Omega}(\vec{R})$ has the elements

$$[\boldsymbol{\Omega}]_{kk'} = \sum_{j=1}^{N} (\vec{e}_k \cdot \vec{\mathfrak{r}}_j^{\,0})\,(\vec{e}_{k'} \cdot \vec{r}_j) \quad , \tag{12}$$

and $\kappa \equiv \kappa(\vec{R})$ is the greatest real root of the equation

$$\kappa^4 - 2\ \mathrm{Tr}(\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega})\,\kappa^2 - 8\,(\det\boldsymbol{\Omega})\,\kappa + \Big(2\ \mathrm{Tr}(\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega}\,\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega}) - [\mathrm{Tr}(\boldsymbol{\Omega}^{\boldsymbol{T}}\boldsymbol{\Omega})]^2\Big) = 0 \quad . \tag{13}$$

When inserted into Eq. (10), the resulting $\boldsymbol{T}(\vec{R})$, i.e. $\boldsymbol{T}(\theta_1, \theta_2, \theta_3)$ expressed in terms of $\vec{R}$, yields $\{q_p\}$ as functions of the position vectors $\{\vec{r}_j\}$. Finally, noting that

$$D_{Lm\mu}(\theta_1, \theta_2, \theta_3) = \left[\frac{2L+1}{8\,\pi^2}\,\frac{(L+m)!\,(L-m)!}{(L+\mu)!\,(L-\mu)!}\right]^{1/2} \times \left(\frac{T_{13} + i\,T_{23}}{2}\right)^{m} \left(\frac{-T_{31} + i\,T_{32}}{1 - T_{33}}\right)^{\mu} P_{L-m}^{m-\mu,m+\mu}(T_{33}) \quad , \tag{14}$$

where $T_{kk'} = [\boldsymbol{T}(\vec{R})]_{kk'}$, completes the prescription for computing $\Psi_{Lmn;\{\nu_p\}}(\vec{R})$ from Eq. (8) for any allowed combination of quantum numbers.

For the three-electron Harmonium atom, described by the Hamiltonian

$$\hat{H} = \frac{1}{2}\sum_{j=1}^{N}(-\hat{\nabla}_j^2 + \omega^2\, r_j^2) + \sum_{j>j'=1}^{N} |\vec{r}_j - \vec{r}_{j'}|^{-1} \tag{15}$$

with $N = 3$, this approach assigns the lowest-energy states to the parent primitive wavefunctions with vanishing vibrational quantum numbers. In particular, the ${}^2P_-$ doublet with $n = 1$ is predicted to be the ground state that lies below the ${}^4P_+$ quartet with $n = 0$ (both states corresponding to the lowest allowed value of $L = 1$) in agreement with the results of

accurate numerical studies [9]. For the four-electron species that forms a Wigner molecule with the equilibrium geometry of a regular tetrahedron, one obtains the $^5S_-$ quintet as the ground state, followed by the $^3P_+$ triplet and the $^1D_+$ singlet as, respectively, the first and second excited states, their parentage being again the primitive functions with vanishing vibrational quantum numbers. This energy ordering matches that obtained from numerical work [10].

Acknowledgements

The research described in this publication has been funded by NCN (Poland) under Grant No. DEC-2012/07/B/ST4/00553. The support from MPI PKS Dresden is also acknowledged.).

References

[1] R. Egger, W. Häusler, C. H. Mak, and H. Grabert, Phys. Rev. Lett. **82**, 3320 (1999).

[2] C. Yannouleas and U. Landman, Phys. Rev. Lett. **82**, 5325 (1999).

[3] J. D. Louck, J. Mol. Spect. **61**, 107 (1976).

[4] E. B. Wilson, Jr., J. C. Decks, and P. C. Cross, *Molecular Vibrations*, Ch. 11, McGraw-Hill, N. York (1955).

[5] J. K. G. Watson, Mol. Phys. **15**, 479 (1968).

[6] W. Kohn, Phys. Rev. **123**, 1242 (1961).

[7] G. W. King, R. M. Hainer, and P. C. Cross, J. Chem. Phys. **11**, 27 (1943).

[8] J. Cioslowski (to be published).

[9] J. Cioslowski, K. Strasburger, and E. Matito, J. Chem. Phys. **136**, 194112 (2012).

[10] J. Cioslowski, K. Strasburger, and E. Matito, J. Chem. Phys. **141**, 044128 (2014).

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An efficient numerical method to solve 2D parabolic convection-diffusion singularly perturbed problems with turning points

C. Clavero[1] and J. Vigo-Aguiar[2]

[1] *IUMA and Department of Applied Mathematics, University of Zaragoza*

[2] *Department of Applied Mathematics, University of Salamanca*

emails: clavero@unizar.es, jvigo@usal.es

Abstract

In this work we analyze, from a numerical point of view, the efficiency of a numerical method used to solve a type of two dimensional parabolic singularly perturbed problems of convection-diffusion type. In the differential equation of the initial and boundary value problem, each component of the convective term has an interior simple turning point, which can be of attractive or repulsive type. The algorithm combines the fractional implicit Euler method, defined on a uniform mesh, to discretize in time, and the classical upwind finite difference scheme, defined on an appropriated mesh, to discretize in space. The fully discrete scheme is uniformly convergent with respect to the diffusion parameter. Some numerical results for different test problems are showed, which corroborate the efficiency and the order of uniform convergence of the numerical method.

Key words: parabolic problem, convection-diffusion, turning point, finite difference scheme, special meshes, uniform convergence

MSC 2000: 65N05, 65N06, 65N10

1 Introduction

In this paper we are interested in the numerical approximation of the solution of a type of 2D parabolic convection-diffusion problems, which is given by the initial and boundary value problem

$$\begin{array}{l} \mathcal{L}u \equiv \dfrac{\partial u}{\partial t} + (\mathcal{L}_{1,\varepsilon}(t) + \mathcal{L}_{2,\varepsilon}(t))\, u = f, \ \text{in } \Omega \times (0,T], \\ u(x,y,0) = \varphi(x,y), \ \text{in } \Omega, \\ u(x,y,t) = 0, \ \text{in } \partial\Omega \times [0,T], \end{array} \tag{1}$$

 ISBN: 978-84-608-6082-2

where $\Omega \equiv (0,1)^2$, and the spatial differential operators $\mathcal{L}_{i,\varepsilon}, i = 1,2$, are defined by

$$\begin{aligned} \mathcal{L}_{1,\varepsilon}(t) &\equiv -\varepsilon \frac{\partial^2}{\partial x^2} + v_1(x,y)\frac{\partial}{\partial x} + k_1(x,y,t), \\ \mathcal{L}_{2,\varepsilon}(t) &\equiv -\varepsilon \frac{\partial^2}{\partial y^2} + v_2(x,y)\frac{\partial}{\partial y} + k_2(x,y,t), \end{aligned} \tag{2}$$

respectively.

We assume that the diffusion parameter ε, $0 < \varepsilon \leq 1$, can be very small, that the reaction terms satisfy $k_i(x,y,t) \geq 0,\ i = 1,2$, and that the convective coefficients are given by

$$v_1(x,y) = -(x-1/2)a_1(x,y),\ v_2(x,y) = -(y-1/2)a_2(x,y), \tag{3}$$

or by

$$v_1(x,y) = (x-1/2)a_1(x,y),\ v_2(x,y) = (y-1/2)a_2(x,y), \tag{4}$$

with a_1 and a_2 such that $a_1(x,y) \geq \alpha_1 > 0,\ a_2(x,y) \geq \alpha_2 > 0$, i.e., we assume that both coefficients of the convective term has a simple turning point at $x = 1/2$ and $y = 1/2$ respectively. We also assume that a_1, a_2, k_1 and k_2 are sufficiently smooth functions, and that sufficient compatibility conditions hold, in order that the exact solution be sufficiently regular. In case (3), the turning points are attractive and therefore only internal layers, located at $x = 1/2$ and/or $y = 1/2$, can appear depending of the smoothness of the right-hand side of the differential equation in (1). On the other hand, in case (4), the turning points are repulsive an therefore, in general, boundary layers at the boundary of the spatial domain can appear in the exact solution.

Henceforth, we denote by N and M the discretization parameters, being the number of mesh intervals in the spatial variables and in time variable, respectively. Below, we always use the pointwise maximum norm, denoted by $\|\cdot\|_D$ (where D is the corresponding domain).

2 The numerical method

In this section we define the fully discrete numerical method that we propose to discretize the continuous problem (1)-(2). The first step is the discretization in time variable. For that, we consider the fractional implicit Euler method, which can be written as a two half step scheme as follows. Let $\tau \equiv T/M$ be the time step, and let us consider the uniform mesh $\bar{I}_M = \{t_n = n\tau,\ n = 0,1,\cdots,M\}$. Let $u^n \approx u(x,y,t_n),\ n = 0,1,\cdots,M$, be the

 ISBN: 978-84-608-6082-2

semidiscrete solutions defined by

$$\begin{array}{l} i) \text{ (initialize)} \\ u^0 = \varphi(x,y), \quad (x,y) \in \Omega. \\ ii) \text{ (first half step)} \\ (I + \tau \mathcal{L}_{1,\varepsilon}(t_{n+1}))u^{n+1/2} = u^n + \tau f_1^{n+1}, \quad (x,y) \in \Omega, \\ u^{n+1/2}(x,y) = 0, \quad (x,y) \in \{0,1\} \times [0,1]. \\ iii) \text{ (second half step)} \\ (I + \tau \mathcal{L}_{2,\varepsilon}(t_{n+1}))u^{n+1} = u^{n+1/2} + \tau f_2^{n+1}, \quad (x,y) \in \Omega, \\ u^{n+1}(x,y) = 0, \quad (x,y) \in [0,1] \times \{0,1\}, \ n = 0, \cdots, M-1, \end{array} \tag{5}$$

with $\mathcal{L}_{1,\varepsilon}(t)$ and $\mathcal{L}_{2,\varepsilon}(t)$ defined in (2) and $f_1^{n+1} = f_1(x,y,t_{n+1}),\ f_2^{n+1} = f_2(x,y,t_{n+1})$.

After the time discretization, to deduce the fully discrete scheme we must discretize in space the one dimensional problems of (5), on a rectangular mesh $\overline{w}_N \equiv I_{1,\varepsilon,N} \times I_{2,\varepsilon,N} \subset \overline{\Omega}$, where

$$I_{1,\varepsilon,N} = \{0 = x_0 < \cdots < x_N = 1\},\ I_{2,\varepsilon,N} = \{0 = y_0 < \cdots < y_N = 1\}.$$

For simplicity, we assume that the number of mesh points is the same at both spatial directions. We denote by $h_{x,i} = x_i - x_{i-1},\ h_{y,i} = y_i - y_{i-1},\ i = 1, \cdots, N$.

Let us denote in the form $\bullet_N$ the discrete functions defined in $\overline{w}_N$, as $[.]_N$ the functions resulting as the operation of restriction of a function defined in $\overline{\Omega}$ to $\overline{w}_N$, and by U_N^n the discrete function which will approach to $[u(x,y,t_n)]_N$ by using the numerical algorithm. On the mesh $\overline{w}_N \times \bar{I}_M$, the fully discrete scheme uses the classical upwind finite difference scheme and it is defined by

$$\begin{array}{l} i)\ U_N^0 = [\varphi(x,y)]_N, \\ ii)\ (I + \tau \mathcal{L}_{1,\varepsilon,N}(t_{n+1}))U_N^{n+1/2} = U_N^n + \tau [f_1^{n+1}]_N(x,y), \quad (x,y) \in w_N, \\ U^{n+1/2}(x,y) = 0, \quad (x,y) \in \{0,1\} \times I_{2,\varepsilon,N}, \\ iii)\ (I + \tau \mathcal{L}_{2,\varepsilon,N}(t_{n+1}))U_N^{n+1} = U_N^{n+1/2} + \tau [f_2^{n+1}]_N(x,y), \quad (x,y) \in w_N, \\ U^{n+1}(x,y) = 0, \quad (x,y) \in I_{1,\varepsilon,N} \times \{0,1\}, \ n = 0,1, \cdots, M-1, \end{array} \tag{6}$$

where $U_N^n \approx u(x,y,t_n),\ (x,y) \in \overline{w}_N,\ n = 0,1,\cdots,M$, and $\mathcal{L}_{1,\varepsilon,N}(t_{n+1}),\ \mathcal{L}_{2,\varepsilon,N}(t_{n+1})$ are given by

$$\begin{array}{l} \mathcal{L}_{1,\varepsilon,N}(t_{n+1})\ W \equiv l_{i-,j} W(x_{i-1},y_j) + l_{ic,j} W(x_i,y_j) + l_{i+,j} W(x_{i+1},y_j),\ i,j = 1,\cdots,N-1, \\ \mathcal{L}_{2,\varepsilon,N}(t_{n+1})\ W \equiv l_{i,j-} W(x_i,y_{j-1}) + l_{i,jc} W(x_i,y_j) + l_{i,j+} W(x_i,y_{j+1}),\ i,j = 1,\cdots,N-1, \end{array} \tag{7}$$

 ISBN: 978-84-608-6082-2

with

$$\begin{aligned}
l_{i-,j} &= \frac{-\varepsilon}{h_{x,i}\tilde{h}_{x,i}} - \frac{(|x_i - 1/2| + (x_i - 1/2))a_1(x_i, y_j)}{2h_{x,i}},\\
l_{i+,j} &= \frac{-\varepsilon}{h_{x,i+1}\tilde{h}_{x,i}} - \frac{(|x_i - 1/2| - (x_i - 1/2))a_1(x_i, y_j)}{2h_{x,i+1}},\\
l_{ic,j} &= k_1(x_i, y_j, t_{n+1}) - l_{i-,j} - l_{i+,j}, \quad i, j = 1, \cdots, N-1,\\
l_{i,j-} &= \frac{-\varepsilon}{h_{y,j}\tilde{h}_{y,j}} - \frac{(|y_j - 1/2| + (y_j - 1/2))a_2(x_i, y_j)}{2h_{y,j}},\\
l_{i,j+} &= \frac{-\varepsilon}{h_{y,j}\tilde{h}_{y,j+1}} - \frac{(|y_j - 1/2| - (y_j - 1/2))a_2(x_i, y_j)}{2h_{y,j+1}}\\
l_{i,jc} &= k_2(x_i, y_j, t_{n+1}) - l_{i,j-} - l_{i,j+}, \quad i, j = 1, \cdots, N-1,
\end{aligned} \tag{8}$$

where $\tilde{h}_{x,i} = (h_{x,i} + h_{x,i+1})/2,\ i = 1, \cdots, N-1$, $\tilde{h}_{y,j} = (h_{y,j} + h_{y,j+1})/2,\ j = 1, \cdots, N-1$.

The definition of meshes $I_{1,\varepsilon,N}, I_{2,\varepsilon,N}$ depends of the case that we consider. In the first one, we assume that (3) holds and that the right-hand side is a continuous function; then, there are not any internal layer and a uniform mesh is sufficient to prove that the fully discrete scheme is uniformly convergent.

In the second case, we assume that (3) holds, but now the right-hand side is discontinuous at the points $(1/2, y),\ 0 \le y \le 1$ and/or at the points $(x, 1/2),\ 0 \le x \le 1$; then, a internal layer appears at $x = 1/2$ and/or $y = 1/2$ which has a size $O(\sqrt{\varepsilon})$. Having into account this behavior of the exact solution, we define a special piecewsise uniform of Shishkin type, which concentrates the mesh points around $x = 1/2$ and $y = 1/2$. We give the details of the construction of $I_{1,\varepsilon,N}$ and analogously we proceed for $I_{2,\varepsilon,N}$.

We define the transition parameter

$$\sigma_x = \min(\frac{1}{4}, m_x\sqrt{\varepsilon}\log N), \tag{9}$$

where m_x a constant to be fixed later. We divide the interval $[0, 1]$ in the subintervals $[0, 1/2 - \sigma_x], [1/2 - \sigma_x, 1/2 + \sigma_x]$ and $[1/2 + \sigma_x, 1]$. Then, the mesh is piecewise uniform having $N/4, N/2$ and $N/4$ subintervals on each one of them. So, the mesh points are given by

$$x_i = \begin{cases} i\dfrac{4(1/2 - \sigma_x)}{N},\ i = 0, \cdots, \dfrac{N}{4}, \\ 1/2 - \sigma_x + (i - \dfrac{N}{4})\dfrac{4\sigma_x}{N},\ i = \dfrac{N}{4} + 1, \cdots, \dfrac{3N}{4}, \\ 1/2 + \sigma_x + (i - \dfrac{3N}{4})\dfrac{4(1/2 - \sigma_x)}{N},\ i = \dfrac{3N}{4} + 1, \cdots, N. \end{cases} \tag{10}$$

Finally, we assume that (4) holds and also that the right-hand side is a continuous function; then a boundary layer appears at the boundary of the spatial domain Ω, which has a size $O(\varepsilon)$. We give the details of the construction of $I_{1,\varepsilon,N}$ and analogously we proceed

for $I_{2,\varepsilon,N}$. We define the transition parameter

$$\sigma_x = \min(\frac{1}{4}, m_x \varepsilon \log N), \tag{11}$$

where m_x a constant to be fixed later. We divide the interval $[0,1]$ in the subintervals $[0,\sigma_x], [\sigma_x, 1-\sigma_x]$ and $[1-\sigma_x, 1]$. Then, the mesh is piecewise uniform having $N/4, N/2$ and $N/4$ subintervals on each one of them. So, the mesh points are given by

$$x_i = \begin{cases} i\dfrac{4\sigma_x}{N},\ i = 0, \cdots, \dfrac{N}{4},\ i = 0, \cdots, \dfrac{N}{4}, \\ \sigma_x + (i - \dfrac{N}{4})\dfrac{2(1-2\sigma_x)}{N},\ i = \dfrac{N}{4}+1, \cdots, \dfrac{3N}{4}, \\ 1 - \sigma_x + (i - \dfrac{3N}{4})\dfrac{4\sigma_x}{N},\ i = \dfrac{3N}{4}+1, \cdots, N. \end{cases} \tag{12}$$

3 Numerical results

In this section we show the numerical results obtained with the algorithm proposed here to solve successfully some problems of type (1). For the shake of simplicity, for all test examples we have chosen the same decomposition for the reaction term, $k_1(x,y,t) = k_2(x,y,t) = k(x,y,t)/2$ and the same decomposition of the right-hand side in the form $f(x,y,t) = f_1(x,y,t) + f_2(x,y,t)$, where $f_2(x,y,t) = f(x,0,t) + y(f(x,1,t) - f(x,0,t))$, $f_1(x,y,t) = f(x,y,t) - f_2(x,y,t)$.

The first example is given by

$$\begin{aligned} &u_t - \varepsilon\Delta u - (x-1/2)(1+xy)u_x - (y-1/2)(x^2+y^2+e^{xy})u_y + (t+\cos(xy))u = \\ &\qquad \pi t^3 e^{-t}\cos(\pi(x+y)),\ (x,y,t) \in \Omega \times [0,1], \\ &u(x,y,t) = 0, (x,y) \text{ in } \partial\Omega \times [0,1], \\ &u(x,y,0) = 0, \text{ in } \Omega. \end{aligned} \tag{13}$$

Figure 1 shows the solution at the final time $t = 1$; from it, we clearly see that there are not any boundary and internal layers.

As the exact solution is unknown, to approximate the maximum pointwise errors, we use a variant of the double-mesh principle. We calculate $\{\hat{u}^N\}$, the numerical solution on the mesh $\{(\hat{x}_i, \hat{y}_j, \hat{t}_n)\}$ containing the original mesh points and its midpoints, i.e.,

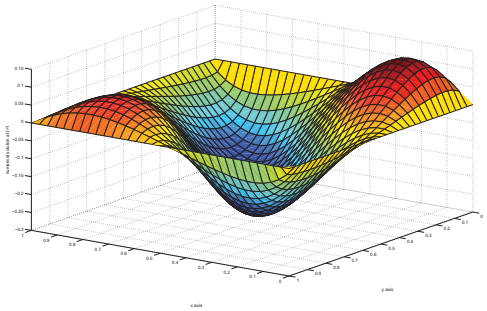
$$\begin{array}{llll} \hat{x}_{2i} = x_i, & i = 0, \ldots, N, & \hat{x}_{2i+1} = (x_i + x_{i+1})/2, & i = 0, \ldots, N-1, \\ \hat{y}_{2j} = y_j, & j = 0, \ldots, N, & \hat{y}_{2j+1} = (y_j + y_{j+1})/2, & j = 0, \ldots, N-1, \\ \hat{t}_{2n} = t_n, & n = 0, \ldots, M, & \hat{t}_{2n+1} = (t_n + t_{n+1})/2, & n = 0, \ldots, M-1. \end{array}$$

Then, the maximum errors at the mesh points of the coarse mesh are approximated by

$$d_{i,j,N,M} = \max_{0 \le n \le M} \max_{0 \le i,j \le N} |u^N(x_i, y_j, t_n) - \hat{u}^N(x_i, y_j, t_n)|, \tag{14}$$

 ISBN: 978-84-608-6082-2

Figure 1: Solution of example (13) for $\varepsilon = 10^{-4}$



and the orders of convergence are given by $q = \log\left(d_{i,j,N,M}/d_{i,j,2N,2M}\right)/\log 2$.

From the double-mesh differences in (14) we obtain the uniform maximum errors by $d^{N,M} = \max_{\varepsilon} d_{i,j,N,M}$, and from them, in a usual way, the corresponding numerical uniform orders of convergence by $q^{uni} = \log\left(d^{N,M}/d^{2N,2M}\right)/\log 2$.

Table 1 displays the results for problem (13) by using our fully discrete scheme on uniform meshes in both space variables. From it, we deduce in this case that on uniform meshes the method is uniformly convergent.

Table 1: Maximum errors and orders of convergence for (13) on uniform meshes

ε	N=16 M=8	N=32 M=16	N=64 M=32	N=128 M=64	N=256 M=128
2^{-6}	1.3866E-2 0.036	1.3528E-2 0.454	9.8770E-3 0.684	6.1490E-3 0.809	3.5098E-3
2^{-8}	1.3443E-2 -.026	1.3685E-2 0.424	1.0202E-2 0.680	6.3698E-3 0.819	3.6108E-3
2^{-10}	1.4088E-2 0.024	1.3860E-2 0.403	1.0485E-2 0.666	6.6078E-3 0.820	3.7440E-3
2^{-12}	1.4266E-2 0.034	1.3935E-2 0.393	1.0611E-2 0.660	6.7169E-3 0.776	3.9221E-3
2^{-14}	1.4311E-2 0.036	1.3957E-2 0.390	1.0649E-2 0.659	6.7463E-3 0.667	4.2501E-3
... 2^{-22}	... ... 1.4326E-2 0.037	... ... 1.3965E-2 0.389	... ... 1.0662E-2 0.658	... ... 6.7576E-3 0.628	... 4.3741E-3
$d^{N,M}$ q^{uni}	1.4326E-2 0.037	1.3965E-2 0.389	1.0662E-2 0.658	6.7576E-3 0.628	4.3741E-3

 ISBN: 978-84-608-6082-2

The second example is the same as in (13), but now the right-hand side is given by

$$f(x,y,t) = \begin{cases} t(2+\sin(x+y)), & \text{if } x \leq 1/2, \\ t^2(1-xy), & \text{if } x > 1/2, \end{cases} \tag{15}$$

for which again the exact solution is unknown. Note that now the right-hand side is discontinuous at the points $(1/2, y)$, $0 \leq y \leq 1$. Figure 2 shows the solution at the final time $t = 1$; from it, we clearly see that there is a internal layer at $y = 1/2$.

Figure 2: Solution of example (13)-(15) for $\varepsilon = 10^{-4}$

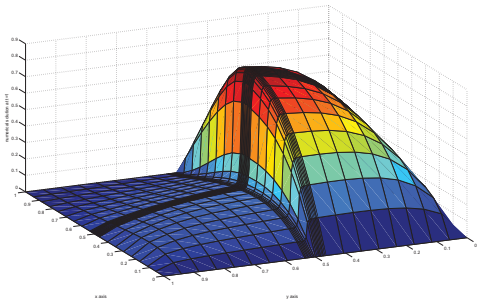


Table 2 displays the results for problem (13)-(15) by using our fully discrete scheme on the corresponding Shishkin mesh, taking $m_x = m_y = 1$ to define the transition parameters σ_x and σ_y in (9). From it, we deduce that the method is an almost first order uniformly convergent method.

The last example is given by

$$\begin{aligned} &u_t - \varepsilon\Delta u + (x-1/2)(1+xy)u_x + (y-1/2)(3-x^2-y^2)u_y + (t+\cos(xy))u = \\ &\qquad\qquad t(xy+\sin(x+y)+2), \ (x,y,t) \in \Omega \times [0,1], \\ &u(x,y,t) = 0, (x,y) \text{ in } \partial\Omega \times [0,1], \\ &u(x,y,0) = 0, \text{ in } \Omega. \end{aligned} \tag{16}$$

Figure 3 shows the solution at the final time $t = 1$; from it, we clearly see that there are boundary layers at the boundary of the spatial domain.

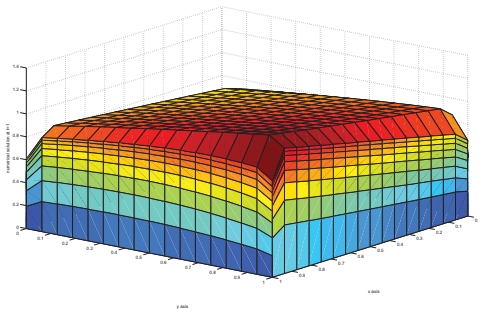
Table 3 displays the results for problem (16) by using our fully discrete scheme on the corresponding Shishkin mesh, taking $m_x = m_y = 1$ to define the transition parameters σ_x and σ_y in (11). From it, we deduce that the method is a first order uniformly convergent method; so, we can conclude that in this example the global errors are dominate by the errors associate to the dicsretization in time.

 ISBN: 978-84-608-6082-2

Table 2: Maximum errors and orders of convergence for (13)-(15) on Shishkin meshes

ε	N=16 M=8	N=32 M=16	N=64 M=32	N=128 M=64	N=256 M=128
2^{-6}	1.0946E-1 0.731	6.5962E-2 0.883	3.5755E-2 0.913	1.8983E-2 0.920	1.0030E-2
2^{-8}	1.2646E-1 0.527	8.7755E-2 0.727	5.3002E-2 0.932	2.7780E-2 0.971	1.4168E-2
2^{-10}	1.2964E-1 0.539	8.9214E-2 0.909	4.7522E-2 0.751	2.8241E-2 0.819	1.6008E-2
2^{-12}	1.3072E-1 0.873	7.1399E-2 0.373	5.5115E-2 0.791	3.1861E-2 0.839	1.7807E-2
2^{-14}	1.3111E-1 0.550	8.9577E-2 0.702	5.5071E-2 0.926	2.8975E-2 0.703	1.7803E-2
... ... 2^{-22}	... ... 1.3138E-1 0.555	... ... 8.9420E-2 0.706	... ... 5.4816E-2 0.791	... ... 3.1691E-2 0.928	... ... 1.6652E-2
$d^{N,M}$ q^{uni}	1.3138E-1 0.553	8.9577E-2 0.701	5.5115E-2 0.791	3.1861E-2 0.839	1.7807E-2

Figure 3: Solution of example (16) for $\varepsilon = 10^{-4}$



Acknowledgements

This work has been partially supported by the project MTM2014-52859 and the Diputación General de Aragón.

References

[1] C. Clavero and F.J. Lisbona, *Uniformly convergent finite difference methods for singularly perturbed problems with turning points*, Numer. Algorithms **4** (1993) 339–359.

Table 3: Maximum errors and orders of convergence for (16) on Shishkin meshes

ε	N=16 M=8	N=32 M=16	N=64 M=32	N=128 M=64	N=256 M=128
2^{-6}	1.1670E-1 0.815	6.6345E-2 0.818	3.7642E-2 0.799	2.1638E-2 0.839	1.2093E-2
2^{-8}	1.2575E-1 0.847	6.9929E-2 0.895	3.7609E-2 0.810	2.1447E-2 0.825	1.2107E-2
2^{-10}	1.3058E-1 0.841	7.2884E-2 0.912	3.8738E-2 0.847	2.1536E-2 0.825	1.2158E-2
2^{-12}	1.3197E-1 0.833	7.4098E-2 0.908	3.9487E-2 0.869	2.1624E-2 0.825	1.2208E-2
2^{-14}	1.3233E-1 0.830	7.4434E-2 0.905	3.9760E-2 0.876	2.1661E-2 0.824	1.2234E-2
...	... ...	... ...	... ...	... ...	...
2^{-22}	1.3245E-1 0.829	7.4548E-2 0.903	3.9859E-2 0.878	2.1682E-2 0.823	1.2252E-2
$d^{N,M}$ q^{uni}	1.3245E-1 0.829	7.4548E-2 0.903	3.9859E-2 0.878	2.1682E-2 0.823	1.2252E-2

[2] C. Clavero, J.L. Gracia, G.I. Shishkin and L.P. Shishkina, *Grid approximation of a singularly perturbed parabolic equation with degenerating convective term and discontinuous right-hand side*, Int. J. Numer. Anal. Mod. **10** (2013) 795–814.

[3] R. Dunne, E. O'Riordan and G.I. Shishkin, *A fitted mesh method for a class of singularly perturbed parabolic problems with a boundary turning point*, Comput. Meth. Appl. Math. **3** (2003) 361–372.

[4] S. Natesan, J. Jayakumar and J. Vigo, *Parameter uniform numerical method for singularly perturbed turning point problems exhibiting boundary layers*, J. Comput. Appl. Math. **158** (2003) 121–134.

[5] H.G. Roos, M. Stynes and L. Tobiska, *Robust numerical methods for singularly perturbed differential equations*, Springer Series in Computational Mathematics, Berlin, 2008.

[6] G.I. Shishkin and L.P. Shishkina, *Difference methods for singular perturbation problems*, Chapman & Hall/CRC Press, Boca Raton, 2009.

[7] R. Vulanovic and P. Farrel, *Continuous and numerical analysis of a multiple boundary turning point problem*, SIAM J. Numer. Anal. **30** (1993) 1400–1418.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Multicriteria design of energy-conscious fuzzy rule-based classifiers for embedded devices

Alberto Cocaña-Fernández[1], José Ranilla[1], Luciano Sánchez[1] and Roberto Gil-Pita[2]

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

[2] *Department of Signal Theory and Communications, University of Alcalá, Spain*

emails: cocanaalberto@gmail.com, ranilla@uniovi.es, luciano@uniovi.es, roberto.gil@uah.es

Abstract

When computationally intensive classifiers are to be implemented in energy constrained embedded devices, there is a tradeoff between accuracy and battery life. Specific machine learning algorithms are needed in order to deploy efficient solutions for wearable devices or medical aids, where the life span of the battery is minimal. In this paper, a design method is proposed for finding rule-based classifiers with a sensible energy consumption for an unbalanced mix of use cases. This method is validated through the design of a last-generation hearing aid, that can self adapt to the environment and detect ambient noise, music or speech without human interaction.

Key words: Energy-conscious machine learning, Fuzzy rule learning, Embedded devices, Hearing aids

1 Introduction: Energy-Efficient Machine Learning

Multiple lines of work within the field of machine learning can be explored in the pursuit of energy-efficient classification systems. For example, a basic approach is to use relatively simple classifiers with low computational costs such as decision trees or rule-based systems. This type of schemes that often exhibit high linguistic interpretability tend to require fewer operations to infer the class for a given instance, thus reducing the energy required to perform the classification. In addition to this, high performing but costly schemes can also be modified to reduce their classification complexity. Examples of this are the approximation

of the Gaussian RBF kernel proposed in [4] to reduce the computational cost of Support Vector Machines (SVM), and the Multiplier-less Artificial Neuron presented in [6] to improve energy consumption on artificial neural networks.

Classifiers can also be built from scratch with complexity in mind. For instance, Multi-objective Genetic Fuzzy Systems jointly maximize classification accuracy and minimise the computational complexity in terms of the number of rules and the size of the antecedents. This is done by producing a set of nondominated fuzzy classifiers with different trade-offs between both objectives through evolutionary multiobjective optimization [2]. Feature selection can also address the aforementioned issue by indirectly reducing the size of the classifier and/or the cost associated to the computation of each feature. Other than this, multi-objective feature selection algorithms exist that reduce the size of the feature subset while maximizing classification performance [8], and multiple simple classifiers can be combined to produce an ensemble of learned models [1].

Despite most classification schemes expend the same computational costs in all inputs, not all instances necessarily require the same classification complexity. As according to [7], a vast majority of inputs can be classified correctly with very low effort, and only a small fraction require the full potential of the learned classifiers. To address this issue, various authors have proposed multi-stage classifiers with increasing levels of complexity [5].

Our study will focus in incremental and hierarchical classifiers, whose code can be short-circuited in part thus different subsets of features are involved when deciding between different pairs of classes. The purpose of the design presented in this work is to restrict the computation of the most expensive features to instances that appear with a lower probability, minimizing the average energy consumption of the classifier.

2 Multiobjective Evolutionary Programming Classifiers

Multiobjective evolutionary algorithms (MOEAs) are to jointly optimise prediction accuracy, feature costs and classification complexity. A population of individuals is iteratively evolved through the application of crossover and mutation operators. The result is a set of non-dominated individuals, also denoted as the Pareto Efficient Frontier.

Each individual in the population is a complete classifier learned according to a predefined grammar. Let us consider a simple multi-rule fuzzy classifier with as many rules as possible classes N_c, each one targeted at determining whether or not the class is the one associated to that rule. Each rule R_n is built according to the following grammar:

RULE $R_n \rightarrow$ if CONDITION then class is C_n

CONDITION $\rightarrow$ (CONDITION $\wedge$ CONDITION)
| (CONDITION $\vee$ CONDITION)
| ASSERTION

ASSERTION $\rightarrow F_x \; is \; \widetilde{Lt} \mid F_x \; is \; \widetilde{Tr} \mid F_x \; is \; \widetilde{Rt}$

where $F_1 \dots F_N$ are the input features of every instance, $\wedge$ is the logical conjunction computed as $(a * b)$, $\vee$ is the logical disjunction computed as $(a + b - a * b)$ and $\widetilde{Lt}$, $\widetilde{Rt}$, $\widetilde{Tr}$ are, respectively, left trapezoidal, right trapezoidal and triangular fuzzy sets [3].

3 Experimental results

Preliminary results in the application of the proposed method to hearing aids support the main hypothesis of this work: rule-based classifiers can be conceived that give up a negligible accuracy while noticeably undercut energy consumption. In Figure 1 a typical Pareto Efficient Frontier is displayed. This graph relates three values: (a) classification cost (b) cost of the features (this is defined as the sum of the energetic costs of the features involved in the classifier) and (c) error rate. In the foreseeable future, a widespread experimentation and a comparison with other machine learning techniques will be provided.

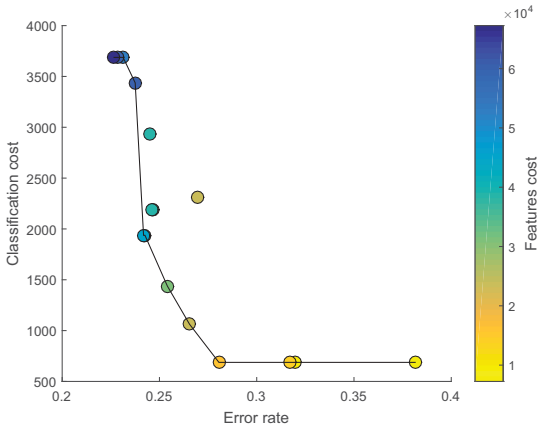


Figure 1: Bidimensional representation of the Pareto Efficient Frontier obtained in the experiment using the test set. Each of the circles represents a fuzzy rule-based classifier with a different tradeoff between energy and accuracy.

4 Concluding remarks and future work

Modern energy constrained embedded devices can support computer intensive algorithms, but there are few studies where machine learning techniques are oriented to maximize the accuracy per watt. Given that the computational cost is instance dependent, the lifespan of the battery will be optimized in terms of the distribution of the expected inputs, combining the highest possible energy gain with the smallest loss of quality. In this respect, the

usefulness of incremental or hierarchical classifiers, where features are only computed when demanded, will be discussed in the full paper.

Acknowledgements

This work has been partially supported by the Ministry of Economy and Competitiveness ("Ministerio de Economía y Competitividad") from Spain/FEDER under grants TEC2015-67387-C4-3-R, TEC2015-67387-C4-4-R and TIN2014-56967-R and by the Regional Ministry of the Principality of Asturias under grant FC-15-GRUPIN14-073.

References

[1] O. Cordon, A. Quirin, and L. Sanchez, *A first study on bagging fuzzy rule-based classification systems with multicriteria genetic selection of the component classifiers*, in 2008 3rd International Workshop on Genetic and Evolving Systems, IEEE, mar 2008, pp. 11–16.

[2] H. Ishibuchi, *Multiobjective Genetic Fuzzy Systems: Review and Future Research Directions*, in 2007 IEEE International Fuzzy Systems Conference, IEEE, jun 2007, pp. 1–6.

[3] H. Ishibuchi, T. Nakashima, and M. Nii, *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining (Advanced Information Processing)*, (2004).

[4] S. Kung and P.-y. Wu, *On efficient learning and classification kernel methods*, in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, mar 2012, pp. 2065–2068.

[5] P. Panda, S. Venkataramani, A. Sengupta, A. Raghunathan, and K. Roy, *Energy-Efficient Object Detection using Semantic Decomposition*, (2015), p. 10.

[6] S. S. Sarwar, S. Venkataramani, A. Raghunathan, and K. Roy, *Multiplier-less Artificial Neurons Exploiting Error Resiliency for Energy-Efficient Neural Computing*, (2016).

[7] S. Venkataramani, A. Raghunathan, J. Liu, and M. Shoaib, *Scalable-effort classifiers for energy-efficient machine learning*, in Proceedings of the 52nd Annual Design Automation Conference on - DAC '15, New York, New York, USA, jun 2015, ACM Press, pp. 1–6.

[8] Z. Yong, G. Dun-wei, and Z. Wan-qiu, *Feature selection of unreliable data using an improved multi-objective PSO algorithm*, Neurocomputing, 171 (2016), pp. 1281–1290.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On a sixth-order family for solving nonlinear models combining derivatives and divided differences

Alicia Cordero[1], Juan R. Torregrosa[1] and María P. Vassileva[2]

[1] *Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, València, Spain*

[2] *Instituto Tecnológico de Santo Domingo (INTEC), Santo Domingo, Dominican Republic*

emails: acordero@mat.upv.es, jrtorre@mat.upv.es, maria.penkova@intec.edu.do

Abstract

Based on a family of bi-parametric iterative methods for solving systems of nonlinear equations, we construct a new family of bi-parametric schemes with sixth-order of convergence. Some elements of the new family have good computational efficiency index, in comparison with other known methods. Some nonlinear systems with arbitrary number of equations are used as numerical tests to contrast with the mentioned known schemes.
Key words: Systems of nonlinear equations, iterative methods, order of convergence.

MSC 2000: AMS codes 65H05, 65H10

1 Introduction

The solution of nonlinear systems is a classic, frequent and important problem for many applications in sciences and engineering. Specifically, we focus our attention to find the solution ξ of a nonlinear system $F(x) = 0$, wherein $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is a sufficiently Frechet differentiable function in an open convex set D. The best known method for solving this kind of problems is Newton's scheme,

$$x^{(k+1)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)}),\ k = 0, 1, 2, \ldots,$$

where $F'(x^{(k)})$ is the Jacobian matrix of function F evaluated in the kth iteration.

Based on Newton's or Newton-like iterations, some higher order methods for computing a solution of nonlinear system $F(x) = 0$ have been proposed in the literature. For example,

 ISBN: 978-84-608-6082-2

among other authors, Montazeri et al. [5] and Hueso et al. [4], developed sixth-order iterative methods requiring two evaluations of function F and two of Jacobian F' per iteration, Sharma and Arora [8] designed a sixth-order method which require three functional and two Jacobian evaluations per iteration. On the other hand, Wang et al. [9] have constructed a seventh-order derivative free iterative method by using the first order divided difference operator $[x, y; F]$ evaluated three times per iteration. The aim of these new schemes is to accelerate the convergence or to improve the computational efficiency.

Hueso et al. in [4] developed several iterative schemes of order six; we use one of them in the numerical section for comparing with our proposed scheme on different test problems. In particular, in [4] the authors present the following method that we denote by HMT

$$\begin{aligned}
y^{(k)} &= x^{(k)} - \frac{2}{3}[F'(x^{(k)})]^{-1}F(x^{(k)}),\\
z^{(k)} &= x^{(k)} - \left[-\frac{1}{2}I + \frac{9}{8}[F'(y^{(k)})]^{-1}F'(x^{(k)}) + \frac{3}{8}[F'(x^{(k)})]^{-1}F'(y^{(k)})\right][F'(x^{(k)})]^{-1}F(x^{(k)}),\\
x^{(k+1)} &= z^{(k)} - \left[-\frac{9}{4}I + \frac{15}{8}[F'(y^{(k)})]^{-1}F'(x^{(k)}) + \frac{11}{8}[F'(x^{(k)})]^{-1}F'(y^{(k)})\right][F'(y^{(k)})]^{-1}F(z^{(k)}).
\end{aligned}$$

On the other hand, Wang et al. in [9] describe the following derivative-free seventh-order scheme, that we denote by WZQT

$$\begin{aligned}
y^{(k)} &= x^{(k)} - B^{-1}F(x^{(k)}),\\
z^{(k)} &= y^{(k)} - \left[3I - 2B^{-1}[y^{(k)}, x^{(k)}; F]\right]B^{-1}F(y^{(k)}),\\
x^{(k+1)} &= z^{(k)} - \left[\frac{13}{4}I - B^{-1}[z^{(k)}, y^{(k)}; F]\left(\frac{7}{2}I - \frac{5}{4}B^{-1}[z^{(k)}, y^{(k)}; F]\right)\right]B^{-1}F(z^{(k)}),
\end{aligned}$$

where $B = \left[x^{(k)} + F(x^{(k)}), x^{(k)} - F(x^{(k)}); F\right]$ and $[x, y; F] : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is the first order divided difference on D.

In order to compare the different methods under the point of view of the computational cost, we recall the computational efficiency index, CI, introduced by the authors in [2], which combine the efficiency index defined by Ostrowski [7] and the number of products-quotients required per iteration. We define this index as $CI = p^{1/(d+op)}$, where p is the order of convergence, d is the number of functional evaluations and op is the number of products-quotients per iteration. Let us remark that for evaluating function F we need n scalar functional evaluations (the coordinate functions of F), whilst for evaluating Jacobian F' it is necessary to evaluate n^2 functions (all the entries of matrix F'). On the other hand, all the iterative methods for solving nonlinear systems require one or more matrix inversion, that is, one or more linear systems must be solved. So, the number of operations needed for solving a linear system plays in this context an important role.

We recall that the number of products and quotients required for solving a linear system by Gaussian elimination is $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$, where n is the size of the system. In addition, for

solving q linear systems, with the same matrix of coefficients, by using LU decomposition we need $\frac{1}{3}n^3+qn^2-\frac{1}{3}n$ products-quotients. By using this information, in Section 3 we compare the computational efficiency index of the different methods used in this manuscript.

In this paper, we design a multidimensional family of bi-parametric iterative schemes, whose local order of convergence is stated in Section 2. Section 3 is devoted to the analysis of the computational efficiency index of some elements of the new family in comparison with other known methods of the same or higher order of convergence. In order to check the theoretical results, some tests on several nonlinear systems of equations are made in Section 4. Finally, some conclusions are presented.

2 Design of the methods

In [1], the authors use a combination of Ostrowski' and Chun's methods for designing a family fourth-order bi-parametric methods

$$x_{k+1} = y_k - \frac{1}{a}\left[\frac{f(x_k)}{f(x_k)+a(b-2)f(y_k)} + \frac{(a-1)f(x_k)+abf(y_k)}{f(x_k)}\right]\frac{f(y_k)}{f'(x_k)}, \tag{1}$$

where y_k is a Newton's step, $y_k = x_k - \frac{f(x_k)}{f'(x_k)}$ and a and b are arbitrary parameters.

Some algebraic manipulations allows us to generalize this family to several variables preserving the local order of convergence four by using the divided difference operator $[\cdot,\cdot;F]:\Omega\times\Omega\subset\mathbb{R}^n\times\mathbb{R}^n\to\mathcal{L}(\mathbb{R}^n)$ defined in [6], such that $[x,y;F](x-y)=F(x)-F(y)$, for any $x,y\in\Omega$. Then, the multidimensional version of the class of iterative methods is

$$\begin{aligned} x^{(k+1)} &= y^{(k)} - G\left(x^{(k)},y^{(k)}\right)\left[F'(x^{(k)})\right]^{-1}F(y^{(k)}),\\ G\left(x^{(k)},y^{(k)}\right) &= \frac{1}{a}\left[(1+ab-2a)I - a(b-2)\left[F'(x^{(k)})\right]^{-1}\left[x^{(k)},y^{(k)};F\right]\right]^{-1}\\ &\quad +\frac{1}{a}\left[(a+ab-1)I - ab\left[F'(x^{(k)})\right]^{-1}\left[x^{(k)},y^{(k)};F\right]\right], \end{aligned} \tag{2}$$

where I is the identity matrix and F' is the jacobian matrix associated with the nonlinear function. Let us remark that we can rewrite function $G\left(x^{(k)},y^{(k)}\right)$ as follows

$$G\left(x^{(k)},y^{(k)}\right) = \left[(1+ab-2a)I - a(b-2)S^{(k)}\right]^{-1}\left[r_1 I - r_2 S^{(k)} + ab(b-2)[S^{(k)}]^2\right], \tag{3}$$

where $r_1 = 3-ab-2a+ab^2$, $r_2 = 2+2ab^2-2a-3ab$ and $S^{(k)} = \left[F'(x^{(k)})\right]^{-1}\left[x^{(k)},y^{(k)};F\right]$.

If we take a particular case of this bi-parametric family of iterative methods for $b=0$ and any non zero values of a, we obtain

$$G\left(x^{(k)},y^{(k)}\right) = \left[(1-2a)I + 2aS^{(k)}\right]^{-1}\left[(3-2a)I - 2(1-a)S^{(k)}\right]. \tag{4}$$

If we represent $2a = 2 - \beta$ we obtain

$$G\left(x^{(k)}, y^{(k)}\right) = \left[(\beta-1)I - (\beta-2)S^{(k)}\right]^{-1}\left[(\beta+1)I - \beta S^{(k)}\right], \tag{5}$$

that is, the extension of the King's family for systems. Now, if $\beta = 2$, function

$$G\left(x^{(k)}, y^{(k)}\right) = 3I - 2S^{(k)} \tag{6}$$

used in (2) is the extension of the Chun's method for systems. However, if $\beta = 0$

$$G\left(x^{(k)}, y^{(k)}\right) = \left[-I + 2S^{(k)}\right]^{-1} \tag{7}$$

and, replacing $G\left(x^{(k)}, y^{(k)}\right)$ in (2), the extension of the Ostrowski's method for systems is obtained.

Among the different elements of multidimensional King's family, the extension of Chun's method is specially efficient in computational terms. So, in order to design a higher order procedure, it will take an special role when specific cases are searched for the predictor step. In general, the multivariate bi-parametric family is used as a predictor, and we add as corrector step in the following form:

$$x^{(k+1)} = z^{(k)} - T(x^{(k)}, y^{(k)}, z^{(k)})\left[F'(x^{(k)})\right]^{-1} F(z^{(k)}), \tag{8}$$

where $z^{(k)}$ is obtained by using (2) and

$$T(x^{(k)}, y^{(k)}, z^{(k)}) = m_1 I + m_2 G(x^{(k)}, y^{(k)}) + m_3 H(y^{(k)}, z^{(k)}), \tag{9}$$

being $G(x^{(k)}, y^{(k)})$ defined in (3) and $H(y^{(k)}, z^{(k)})$ is calculated as

$$H(y^{(k)}, z^{(k)}) = (n_1 + n_2)I - n_2 G(x^{(k)}, y^{(k)})\left[F'(x^{(k)})\right]^{-1}[y^{(k)}, z^{(k)}; F]. \tag{10}$$

Theorem 1 *Let $F : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ sufficiently differentiable at Ω and let $\xi \in \Omega$ the solution of system of nonlinear equations $F(x) = 0$. Let us assume that $F'(x)$ is continuous and nonsingular in ξ. We consider that $x^{(0)}$ is a initial approximation close enough to ξ. Then, the sequence $\{x^{(k)}\}_{k\geq 0}$ obtained by using the iterative scheme*

$$\begin{aligned} y^{(k)} &= x^{(k)} - \left[F'(x^{(k)})\right]^{-1} F(x^{(k)}), \\ z^{(k)} &= y^{(k)} - G(x^{(k)}, y^{(k)})\left[F'(x^{(k)})\right]^{-1} F(y^{(k)}), \\ x^{(k+1)} &= z^{(k)} - T(x^{(k)}, y^{(k)}, z^{(k)})\left[F'(x^{(k)})\right]^{-1} F(z^{(k)}), \end{aligned} \tag{11}$$

converges to ξ with order of convergence six if $m_1 = -n_1 m_3$ and $m_2 = 1$ for any $n_2 \neq 0$, n_1 and m_3. Here, $G(x^{(k)}, y^{(k)})$ is according to (3), $T(x^{(k)}, y^{(k)}, z^{(k)})$ according to (9) and $H(y^{(k)}, z^{(k)})$ according to (10).

By replacing in (9) the conditions on parameters m_1, m_2 and n_2 that guarantee the order of convergence, we obtain the following expression of $T(x^{(k)}, y^{(k)}, z^{(k)})$,

$$T(x^{(k)}, y^{(k)}, z^{(k)}) = (3 + n_2 m_3)I - 2G(x^{(k)}, y^{(k)}) - n_2 m_3 G(x^{(k)}, y^{(k)}) R(x^{(k)}, y^{(k)}, z^{(k)}), \quad (12)$$

where $R(x^{(k)}, y^{(k)}, z^{(k)}) = \left[F'(x^{(k)})\right]^{-1} [y^{(k)}, z^{(k)}; F]$. We note that this function contains two divided difference operators, involving a high computational cost for solving nonlinear systems. By choosing $m_3 = 0$, we eliminate this problem and $T(x^{(k)}, y^{(k)}, z^{(k)})$ takes the form:

$$T(x^{(k)}, y^{(k)}, z^{(k)}) = 3I - 2S, \quad (13)$$

that corresponds to the weight function involving Chun's method at the second step. Therefore, the additional computational cost is minimum by using this expression of T at third step of the iterative process and the same weight function at the second step. In the following, we denote by CTV this sixth-order method.

3 Computational efficiency

Our aim in this section is to compare the efficiency of the methods by using the computational efficiency index CI, presented in the Introduction. Let us recall that $CI = p^{1/(d+op)}$ where p is the order of convergence, d is the number of functional evaluations per iteration and op is the number of products-quotients per iteration. To compute F in any iterative method we need to calculate n scalar functions. The number of scalar functional evaluations is n^2 for any new derivative F'. When we compute a first-order divided difference $[x, y; F]$ we need $n(n-1)$ scalar functional evaluations, where $F(x)$ and $F(y)$ are computed separately.

On the other hand, in order to compute an inverse linear operator we solve a $n \times n$ linear system where we have to do $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ products-quotients for obtaining LU decomposition and solving two triangular linear systems. In addition, we need n^2 divisions for computing a first order divided difference, n products for scalar product and n^2 products for matrix-vector multiplication.

Taking into account the previous considerations, we give the index CI of the method CTV. For each iteration we need to evaluate function F three times, once the Jacobian F' and once the divided difference operator, so $2n^2 + 2n$ functional evaluations are needed. In addition, we must to solve five linear systems with $F'(x^{(k)})$ as coefficient matrix (that is $(1/3)n^3 + 3n^2 - (1/3)n$ products-quotients), two matrix-vector products (n^2 products-quotients) and n^2 divisions in the construction of the divided difference operator. Therefore, the value of index CI for method CTV on a nonlinear system of size $n \times n$ is

$$CI_{CTV} = 6^{\frac{1}{\frac{1}{3}n^3 + 9n^2 + \frac{5}{3}n}}.$$

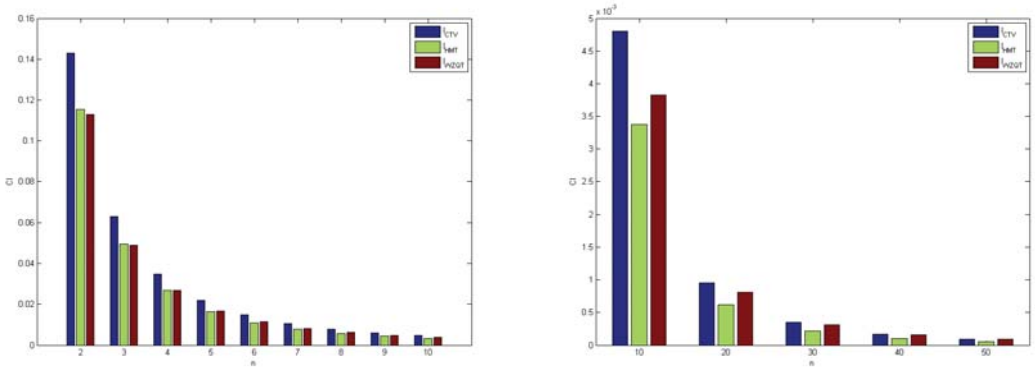
In Table 1 we show index CI of the schemes CTV, HMT, WZQT. In it, NFE is the number of functional evaluations, $NLS1$ denotes the number of linear systems solved with the same matrix of coefficients, $NLS2$ is the number of linear systems solved with other matrix of coefficient, $M \times V$ denotes the number of matrix-vector products and DD the number of divided difference operators.

Method	NFE	$NLS1$	$NLS2$	DD	$M \times V$	CI
CTV	$2n^2+2n$	5	0	1	2	$6^{\frac{1}{\frac{1}{3}n^3+9n^2+\frac{5}{3}n}}$
HMT	$2n^2+2n$	3	3	0	3	$6^{\frac{1}{\frac{2}{3}n^3+11n^2+\frac{4}{3}n}}$
WZQT	$3n^2$	6	0	3	3	$7^{\frac{1}{\frac{1}{3}n^3+15n^2-\frac{1}{3}n}}$

Table 1: Functional evaluations and products-quotients of the methods

Let us observe that method WZQT needs to solve, per iteration, six linear systems with coefficients matrix B, to evaluate three divided differences, that is $3n^2$ quotients, to make three products matrix-vector and also $3n^2$ functional evaluations corresponding to three evaluations of F and three evaluations of the divided difference $[x, y; F]$. However, its order of convergence is seven.

In Figure 1, we can observe the value of index CI for the compared methods and different sizes of the nonlinear system. Let us remark that, for all the analyzed cases, the



(a) Size from 2 to 10 (b) Size from 10 to 50

Figure 1: Index CI for different sizes of the system

best index CI corresponds to method CTV.

4 Numerical results

To illustrate the convergence behavior of the proposed method CTV, we consider some numerical examples and compare the performance with the schemes described in the Introduction and denoted by HMT and WZQT. All computations are performed in the programming package Matlab $R2013b$ using variable precision arithmetics with 2000 digits of mantissa. The divided differences performed are of order 1. For every method, we analyze the number of iterations (iter) needed to converge to the solution such that $\|x^{(k+1)} - x^{(k)}\| < 10^{-100}$ and $\|F(x^{(k+1)})\| < 10^{-100}$ is satisfied, where $\|\cdot\|$ denotes the Euclidean norm. To verify the theoretical order of convergence (p), we calculate the computational order of convergence (ACOC) introduced in [3] as

$$p \approx ACOC = \frac{\ln\left(\|x^{(k+1)} - x^{(k)}\|/\|x^{(k)} - x^{(k-1)}\|\right)}{\ln\left(\|x^{(k)} - x^{(k-1)}\|/\|x^{(k-1)} - x^{(k-2)}\|\right)}.$$

Let us remark that, if the entries of vector ACOC do not stabilize their values along the iterative process, it is marked as '-'.

In Tables 2 to 4, we show the numerical results obtained applying the described methods on three academic nonlinear systems appearing in [9].

Example 1: The first nonlinear system to be solved is defined by $F_1(x_1, x_2, \ldots, x_n)$, whose coordinate functions are $f_1^i = x_i^2 x_{i+1} - 1$, for $i = 1, 2, \ldots, n-1$ and $f_1^n = x_n^2 x_1 - 1$. Moreover, the searched root is $\xi \approx (1, 1, \ldots, 1)^T$.

As this problem has undefined size, we choose $n = 100$, in order to check how the methods behave with a big-sized system. In Table 2, the obtained results confirm the theoretical order of convergence and show the good performance of the three methods.

Methods	iter	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	$ACOC$
CTV	5	2.12e-595	0.0	5.9997
HMT	5	1.27e-875	0.0	5.9999
WZQT	5	2.35e-954	0.0	7.0000

Table 2: Example 1, $x^{(0)} = (1.25, 1.25, \ldots, 1.25)^T$, with size $n = 100$

Example 2: Also used in [9], this nonlinear system has undetermined size, being the involved function $F_2(x_1, x_2, \ldots, x_n)$, whose coordinate functions are defined by $f_2^i = x_i - \cos\left(2x_i - \sum_{i=1}^4 x_i\right)$, $i = 1, 2, \ldots, n-1$, being in this case $\xi \approx (0.5149, 0.5149, \ldots, 0.5149)^T$.

In this case, we have selected $n = 4$ to make the numerical tests, whose results can be seen at Table 3. One of the known methods, HMT, has not reached convergence in a maximum of 50 iterations and it is marked in Table 3 as 'nc'. The convergence of other

two methods to the root is reached but, the instability of the process in both methods is reflected in the calculated ACOC, that is oscillating until the two last iterates.

Methods	iter	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	$ACOC$
CTV	18	8.87e-1034	2.7e-2008	-
HMT	nc	-	-	-
WZQT	19	8.14e-424	1.35e-2008	-

Table 3: Example 2, $x^{(0)} = (-1, -1, \ldots, -1)^T$, with size $n = 4$

Example 3: The last problem to be solved is defined by the nonlinear function $F_3(x_1, x_2, \ldots, x_n)$ whose coordinate functions are $f_3^i = \sum_{j=1, j\neq i}^{n} x_j - \exp(-x_i)$, $i = 1, 2, \ldots, n$, being in this case the root $\xi \approx (0.07714, 0.07714, \ldots, 0.07714)^T$, whose performance is presented in Table 4, where it can be observed that the seventh-order method does not converge after 50 iterations and proposed scheme CTV and known method HMT show similar results, in terms of precision and stability.

Methods	iter	$\|x^{(k+1)} - x^{(k)}\|$	$\|F(x^{(k+1)})\|$	$ACOC$
CTV	4	1.26e-377	1.88e-2007	5.9639
HMT	4	1.34e-353	1.93e-2007	5.9797
WZQT	nc	-	-	-

Table 4: Example 3, $x^{(0)} = (0.55, 0.55, \ldots, 0.55)^T$, with size $n = 50$

5 Conclusions

By summarizing, we have defined a sixth-order class of iterative methods for solving systems of nonlinear equations, on the basis of the best element of a known 4th-order family. From the numerical point of view, the simplest element of the family has been selected and tested numerically among other known schemes of same and higher order of convergence. Its good performance has been showed, even when the other methods fail.

Acknowledgements

This research was partially supported by Ministerio de Economía y Competitividad MTM2014-52016-C02-2-P and FONDOCYT 2014-1C1-088 República Dominicana.

References

[1] A. Cordero, J. G. Maimó, J. R. Torregrosa and M. P. Vassileva, *Solving nonlinear problems by Ostrowski-Chun type parametric families*, J. Math. Chem. **53** (2015) 430–449.

[2] A. Cordero, J.L. Hueso, E. Martínez and J.R. Torregrosa, *A modified Newton-Jarratt's composition*, Numer. Algor. **55** (2010) 87–99.

[3] A. Cordero and J.R. Torregrosa, *Variants of Newton's method using fifth-order quadrature formulas*, Appl. Math. Comput. **190** (2007) 686–698.

[4] J. L. Hueso, E. Martínez and C. Teruel, *Convergence, effiency and dinamimics of new fourth and sixth order families of iterative methods for nonlinear systems*, J. Comput. Appl. Math. **275** (2015) 412–420.

[5] H. Montazeri, F. Soleymani, S. Shateyi and S. S. Motsa, *On a new method for computing the numerical solution of systems of nonlinear equations*,J. Appl. Math. **2012** (2012) ID. 751975, 15 pages.

[6] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic, New York, 1970.

[7] A.M. Ostrowski, *Solution of equations and systems of equations*, Prentice-Hall, Englewood Cliffs, New York, 1964.

[8] J. R. Sharma, H. Arora, *Efficient Jarratt-like methods for solving systems of nonlinear equations*, Calcolo **51** (2014) 193–210.

[9] X. Wang, T. Zhang, W. Qian and M. Teng, *Seventh-order derivative-free iterative method for solving nonlinear systems*, Numer. Algor. **70** (2015) 545–558.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On the dynamics of a class of iterative methods with memory for solving nonlinear equations

Alicia Cordero[1] **and Juan R. Torregrosa**[1]

[1] *Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, València, Spain*

emails: acordero@mat.upv.es, jrtorre@mat.upv.es

Abstract

The dynamical analysis of iterative methods without memory for solving nonlinear equations, by using complex dynamics tools, is a very useful technique to study their stability and reliability. Nevertheless, this technique can not be used on iterative schemes with memory and it is necessary to define it as a real multidimensional discrete dynamical system in order to afford this task. Then, the stability of the fixed points of its rational operator associated on quadratic polynomials can be studied. A parametric family of order four, applied on quadratic polynomials, is studied, showing the bifurcations diagrams and the appearance of chaos.

Key words: Nonlinear equations, iterative method with memory, basin of attraction, stability, bifurcation.

1 Introduction

Our goal in this paper is to carry out a dynamical study of the iterative methods with memory designed for solving a nonlinear equation $f(x) = 0$. As the fixed point iteration function has more than one variable, an auxiliary function is introduced to facilitate the calculations. Moreover, specific dynamical concepts are adapted to achieve the appropriate numerical sense.

On the other hand, we also analyze the local convergence of the method with memory under study. To get this aim, we use the following result, that can be found in [6].

 ISBN: 978-84-608-6082-2

Theorem 1 *Let ψ be an iterative method with memory that generates a sequence $\{x_k\}$ of approximations to the zero α of function $f(x)$, and let this sequence converges to α. If there exist a nonzero constant η and nonnegative numbers t_i, $i = 0, 1, \ldots, m$, such that the inequality*

$$|e_{k+1}| \leq \eta \prod_{i=0}^{m} |e_{k-i}|^{t_i}$$

holds, then the R-order of convergence of the iterative method ψ satisfies the inequality

$$O_R(\psi, \alpha) \geq s^*,$$

where s^* is the unique positive root of the equation

$$s^{m+1} - \sum_{i=0}^{m} t_i s^{m-i} = 0.$$

1.1 Iterative methods with memory as discrete dynamical systems

By using the same structure defined in [1] by the authors, the expression of an iterative method with memory, which uses two previous iterations to calculate the following estimation, is

$$x_{k+1} = g(x_{k-1}, x_k), \quad k \geq 1,$$

where x_0 and x_1 are the initial estimations. A fixed point of this method will be obtained when not only $x_{k+1} = x_k$, but also $x_{k-1} = x_k$. In order to obtain them, we define the *fixed point function* $G : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ by means of:

$$\begin{aligned} G\left(x_{k-1}, x_k\right) &= (x_k, x_{k+1}), \\ &= (x_k, g(x_{k-1}, x_k)), \quad k = 1, 2, \ldots, \end{aligned}$$

being x_0 and x_1 the initial estimations. This definition can be extended in a natural way to adapt it to iterative schemes with memory that use more than two previous iterations per step.

In the following we recall some basic dynamical concepts.

Definition 1 *If a fixed point (z, x) of operator G is different from (r, r), where r is a zero of f, it is called* strange fixed point.

Definition 2 *Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be a vectorial function. The orbit of a point $\bar{x} \in \mathbb{R}^2$ is defined as the set of successive images of $\bar{x}$ by the vector function, $\{\bar{x}, G(\bar{x}), \ldots, G^m(\bar{x}), \ldots\}$.*

Definition 3 *A point $x^* \in \mathbb{R}^2$ is a k-periodic point if $G^k\left(x^*\right) = x^*$ and $G^p\left(x^*\right) \neq x^*$, for $p = 1, 2, \ldots, k-1$.*

 ISBN: 978-84-608-6082-2

The dynamical behavior of the orbit of a point of $\mathbb{R}^2$ is classified depending on its asymptotical behavior. The stability of fixed points for multivariable nonlinear operators, see for example [7], satisfies the following statements:

Theorem 2 *Let G from $\mathbb{R}^n$ to $\mathbb{R}^n$ be $\mathcal{C}^2$. Assume x^* is a k-periodic point. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $G'(x^*)$.*

a) If all the eigenvalues λ_j have $|\lambda_j| < 1$, then x^* is attracting.

b) If one eigenvalue λ_{j_0} has $|\lambda_{j_0}| > 1$, then x^* is unstable, that is, repelling or saddle.

c) If all the eigenvalues λ_j have $|\lambda_j| > 1$, then x^* is repelling.

In addition, a fixed point is called *hyperbolic* if all the eigenvalues λ_j of $G'(x^*)$ have $|\lambda_j| \neq 1$. In particular, if there exist an eigenvalue λ_i such that $|\lambda_i| < 1$ and an eigenvalue λ_j such that $|\lambda_j| > 1$, the hyperbolic point is called *saddle point*.

Moreover, a point x is a *critical point* of G if the associate Jacobian matrix $G'(x)$ satisfies $det(G'(x)) = 0$. One particular case of critical points, for iterative methods of convergence order higher than two, are those fixed points with null eigenvalues $\lambda_j = 0$, $\forall j$. These points are called *superattracting*, as an extension of the scalar case.

Then, if x^* is an attracting fixed point of function G, its *basin of attraction* $\mathcal{A}(x^*)$ is defined as the set of pre-images of any order such that

$$\mathcal{A}(x^*) = \left\{ x^{(0)} \in \mathbb{R}^n : G^m(x^{(0)}) \to x^*, m \to \infty \right\}. \tag{1}$$

The set of the different basins of attraction define the *dynamical plane* of the system. The dynamical plane of a method is built by iterating a mesh of points and painting them in different colors depending on the attractor they converge to.

Summarizing, in this paper we transform a family of iterative methods without memory in a class of schemes with memory holding the derivatives, stating its local order of convergence in Section 2. Later on, the real multivatiate tools described in the Introduction allows us to analyze in Section 3 the stability of the family, showing wide areas of good performance but also undesired elements, as chaotic attractors.

2 Modified family with memory

The fourth-order family of parametric methods without memory under study was presented in [4] as an efficient class to estimate the solution of nonlinear systems of equations. Its iterative expression in the scalar case is

$$y_k = x_k - \theta \frac{f(x_k)}{f'(x_k)},$$

 ISBN: 978-84-608-6082-2

$$\begin{aligned} t_k &= x_k - \frac{f(y_k)+\theta f(x_k)}{f'(x_k)}, \\ x_{k+1} &= x_k - \frac{f(t_k)+f(y_k)+\theta f(x_k)}{f'(x_k)}, \quad k=1,2,\ldots \end{aligned}$$

and its local order of convergence is three, being fourth-order for $\theta = \pm 1$, under standard conditions. We denote this family by HMT.

In [1] this family was modified, by substituting the derivative $f'(x_k)$ appearing in all the steps by $f[x_k, w_k]$ (where $w_k = x_k + \gamma_k f(x_k)$), to obtain a class of methods with memory, denoted by MHMT. Its iterative expression is

$$\begin{aligned} \gamma_k &= -\frac{2}{f[x_k, x_{k-1}]}, \\ w_k &= x_k + \gamma_k f(x_k), \\ y_k &= x_k - \theta \frac{f(x_k)}{f[x_k, w_k]}, \\ t_k &= x_k - \frac{f(y_k)+\theta f(x_k)}{f[x_k, w_k]}, \\ x_{k+1} &= x_k - \frac{f(t_k)+f(y_k)+\theta f(x_k)}{f[x_k, w_k]}, \quad k=1,2,\ldots \end{aligned} \tag{2}$$

and its order of convergence was proved to be at least $\frac{1}{2}\left(3+\sqrt{13}\right)$ if $\theta \neq 0$, and $2+\sqrt{6}$ if $\theta = 1$. Its stability properties were studied, by using some tools of multidimensional real dynamics, and although it was showed to be stable in general, some unstable elements were found; in particular, two strange attractors were detected.

2.1 Design and local convergence

It is possible to design other classes with memory from HMT family, preserving the derivatives. In this paper, we add to the derivative appearing in each step the term $\gamma f(x_k)$, being the accelerating parameter γ the same at each step of the iterative procedure. The iterative expression of the resulting class is

$$\begin{aligned} y_k &= x_k - \theta \frac{f(x_k)}{f'(x_k)+\gamma f(x_k)}, \\ t_k &= x_k - \frac{f(y_k)+\theta f(x_k)}{f'(x_k)+\gamma f(x_k)}, \\ x_{k+1} &= x_k - \frac{f(t_k)+f(y_k)+\theta f(x_k)}{f'(x_k)+\gamma f(x_k)}, \quad k=1,2,\ldots \end{aligned}$$

It is easy to prove that the third-order of convergence is held and the error equation

is $e_{k+1} = (1-\theta)(\gamma+2c_2)(\gamma+(1+\theta)c_2)e_k^3 + O(e_k^4)$, where $c_2 = \dfrac{f''(\alpha)}{2f'(\alpha)}$. Moreover, we get fourth-order of convergence for $\theta = 1$ (as in the original method), but not for $\theta = -1$.

To transform the resulting iterative family in other one with memory increasing the order of convergence it is enough to define $\gamma = -2c_2$; however, it has no sense to use the zero α. We propose the following estimation by using only divided differences,

$$\gamma_k = -\frac{f'[x_k, x_{k-1}]}{f[x_k, x_{k-1}]}.$$

So, we get

$$\begin{aligned}
\gamma_k &= -\frac{f'(x_k)-f'(x_{k-1})}{f(x_k)-f(x_{k-1})} \\
y_k &= x_k - \theta\frac{f(x_k)}{f'(x_k)+\gamma_k f(x_k)}, \\
t_k &= x_k - \frac{f(y_k)+\theta f(x_k)}{f'(x_k)+\gamma_k f(x_k)}, \\
x_{k+1} &= x_k - \frac{f(t_k)+f(y_k)+\theta f(x_k)}{f'(x_k)+\gamma_k f(x_k)}, \quad k = 1, 2, \ldots
\end{aligned} \tag{3}$$

denoted by MF. The local convergence of this scheme is analyzed in the following result.

Theorem 3 *Let α be a simple zero of a sufficiently differentiable function $f : D \subset \mathbb{R} \to \mathbb{R}$ in an open interval D. If x_0 and x_1 are sufficiently close to α, then the order of convergence of method with memory (3) is at least $\frac{1}{2}\left(3+\sqrt{13}\right)$. The error equation is*

$$e_{k+1} = -(\theta-1)^2 c_2(2c_2^2 - 3c_3)e_{k-1}e_k^3 + O_4(e_{k-1}e_k),$$

where $c_j = \dfrac{1}{j!}\dfrac{f^{(j)}(\alpha)}{f'(\alpha)}$, $j = 2, 3, \ldots$ However, if $\theta = 1$, the error equation is

$$e_{k+1} = (-4c_2^5 + 12c_2^3c_3 - 9c_2c_3^2)e_{k-1}^2e_k^4 + O_6(e_{k-1}e_k),$$

being the local order $2+\sqrt{6}$.

As our aim is to analyze the dynamical behavior of the proposed families on real quadratic polynomials, we will study the fixed point operator associated to the presented families on $p_1(x) = x^2-1$, $p_2(x) = x^2+1$ and $p_3(x) = x^2$, that will be denoted by $M_{1,\theta}(z,x)$, $M_{2,\theta}(z,x)$ and $M_{3,\theta}(z,x)$, respectively.

In order to analyze the stability of the members of the class MF, we study the asymptotic behavior of the fixed points of the respective rational functions obtained by applying the fixed point operator on each one of the polynomials, $p_i(x)$, $i = 1, 2, 3$, respectively.

 ISBN: 978-84-608-6082-2

3 Multidimensional dynamical analysis

In this section, we analyze the dynamics of the operator associated to family MF on quadratic polynomials. The associate fixed point operator on $p_1(x) = x^2 - 1$ is

$$M_{1,\theta}(z,x) = \left(x, x - A - B - (x+z)\frac{(x-A-B)^2-1}{2(1+xz)}\right),$$

where

$$A = \frac{\left(x^2-1\right)(x+z)(\theta(x-1)(x+z)-2(1+xz))(\theta(x+1)(x+z)-2(1+xz))}{8(xz+1)^3}$$

and

$$B = \frac{\theta\left(x^2-1\right)(x+z)}{2(xz+1)}.$$

Let us observe that the previous operator depends on three elements the last iteration, x_k (denoted by x), the previous one x_{k-1} denoted by z and one free ($\theta \neq 0$) parameter, θ.

By the way the bidimensional operator $M_{1,\theta}(z,x)$ is constructed, it has a fixed point at (z,x) if $M_{1,\theta}(z,x) = (z,x)$. It means that $z = x$ and then, all the fixed points have two equal components. In the following result we present the different fixed points and their respective stability.

Proposition 1 *The fixed points (and their stability) of the operator associated to MF on quadratic polynomial $p_1(x)$ are:*

a) Points $(1,1)$ and $(-1,-1)$ associated to the roots, being both superattracting.

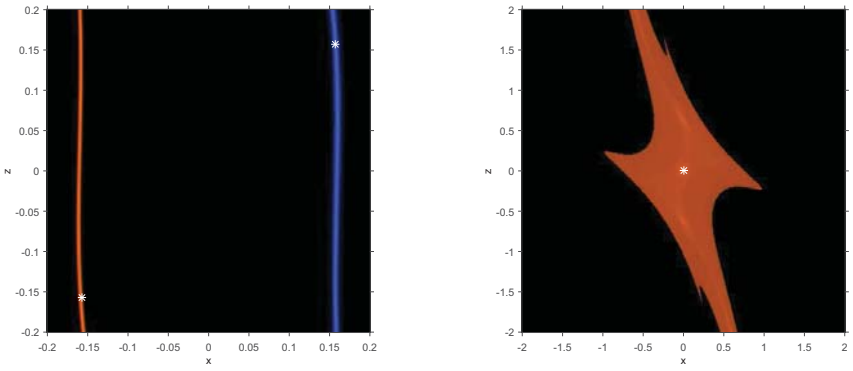
b) The origin $(z,x) = (0,0)$, which is an attracting fixed point for $-4 < \theta < -2$, it is repulsive if $a < -4$ and it is a saddle point if $a > -2$.

c) The real roots of polynomial

$$\begin{aligned} m(x) = {} & 2+\theta+\left(9-2\theta^2\right)x^2+\left(17-3\theta+2\theta^2+2\theta^3\right)x^4+\left(18+4\theta^2-4\theta^3-\theta^4\right)x^6 \\ & +\left(12+3\theta-4\theta^2+3\theta^4\right)x^8+\left(5-2\theta^2+4\theta^3-3\theta^4\right)x^{10} \\ & +\left(1-\theta+2\theta^2-2\theta^3+\theta^4\right)x^{12}, \end{aligned}$$

whose number varies depending on the range of parameter θ: there are two real saddle points if $\theta < -2$, none if $-2 \leq \theta < 6.66633$, two non-hyperbolic points if $\theta \approx 6.66633$ and four (two saddle and two repulsive points) if $\theta > 6.66633$.

Although the fixed point function is different from that analyzed in [1], this result coincides exactly with the one obtained in that manuscript for the family (2), we refer to that reference for the proof.



(a) $\theta = -1.98$ (b) $\theta = -3$

Figure 1: Dynamical plane of MF method on $p_2(x) = x^2 + 1$

Now, we study the behavior of the rational function associated to family MF on $p_2(x) = x^2 + 1$

$$M_{2,\theta}(z,x) = \left(x, C - (x+z)\frac{1+C^2}{2(xz-1)} - (x+z)\frac{\left(C - \frac{(1+C^2)}{2(xz-1)}\right)^2 + 1}{2(xz-1)} \right),$$

where

$$C = x + \frac{\theta\left(x^2+1\right)(x+z)}{2(1-xz)}.$$

Proposition 2 *The fixed points (and their stability) of the operator $M_{2,\theta}(z,x)$ are:*

a) The origin $(z,x) = (0,0)$, which is an attracting fixed point for $-4 < \theta < -2$, it is repulsive if $\theta < -4$ and it is a saddle point if $\theta > -2$.

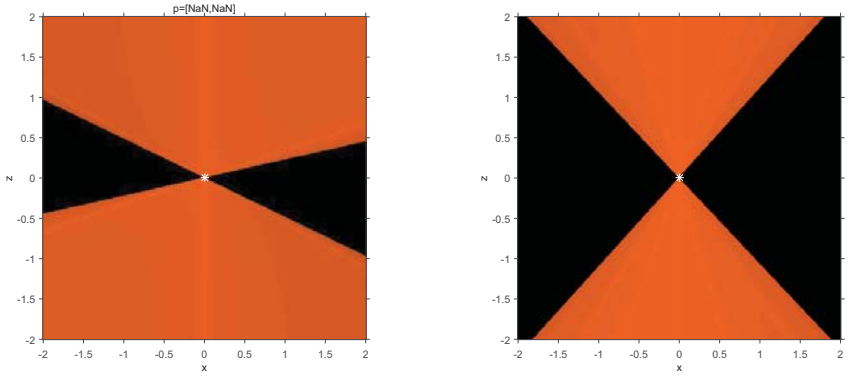
c) The real roots of polynomial

$$\begin{aligned} r(x) = {} & 2 + \theta - \left(9 - 2\theta^2\right)x^2 + \left(17 - 3\theta + 2\theta^2 + 2\theta^3\right)x^4 - \left(18 + 4\theta^2 - 4\theta^3 - \theta^4\right)x^6 \\ & + \left(12 + 3\theta - 4\theta^2 + 3\theta^4\right)x^8 - \left(5 - 2\theta^2 + 4\theta^3 - 3\theta^4\right)x^{10} \\ & + \left(1 - \theta + 2\theta^2 - 2\theta^3 + \theta^4\right)x^{12}, \end{aligned}$$

whose number varies depending on the range of parameter θ: there are two real saddle points if $\theta < -2$, two non-hyperbolic points if $\theta = -2$, four (two saddle and two attracting fixed points) if $-2 < \theta < -1.97095$, two non-hyperbolic points if $\theta \approx -1.97095$ and none if $\theta > -1.97095$.

 ISBN: 978-84-608-6082-2

Some of the behavior shown in the previous proposition is visualized in Figure 1. Specifically, Figure 1a shows the two small basins of attraction of the simultaneously attracting strange fixed points for $\theta = -1.98$. When $\theta = -3$, $(0, 0)$ is the only attracting fixed point, as it can be observed in Figure 1b.



(a) $\theta = 0.5$ (b) $\theta = 1.5$

Figure 2: Dynamical plane of MF method on $p_3(x) = x^2$

Finally, the rational function associated to family MF on $p_3(x) = x^2$ is

$$M_{3,\theta}(z, x) = \left(x, -\frac{x(\theta(x+z) - 2z)\left(\theta(x+z)^2 + 2z(z-x)\right) D}{128 z^7} \right),$$

where $D = \theta^2(x+z)^4 - 4\theta xz(x+z)^2 + 4z^2\left(x^2 + 3z^2\right)$.

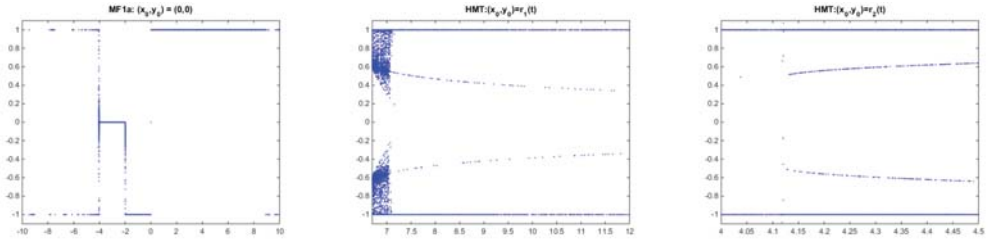
Let us remark that the only fixed point of the operator associated to MF on quadratic polynomial $p_3(x)$ is $(z, x) = (0, 0)$. In this case, an indetermination appears when its stability is analyzed, so it cannot be determined by using Robinson's theorem. However, the dynamical planes in Figure 2 show that it behaves as a saddle point and it remains at the Julia set.

In order to deep in the analysis of the whole parametric family, it is usual to plot a bifurcation diagram, called Feigenbaum diagram, to study the changes of behavior of a fixed point, depending on the values of parameter θ. These kind of diagrams show us the way from regularity to chaos, if it exists (see, for example, [2], [3], [5]).

3.1 Bifurcation diagrams

Now we present the bifurcation diagrams of the map associated to MF family on quadratic polynomials $p_i(x)$ $i = 1, 2, 3$, by using as a starting point each one of the strange fixed points

of the map and observing the ranges of the parameter θ where changes of stability or other behavior happen. This allows us to check the studied dependence of the stability of these points on the parameter and also to find numerically strange attractors, if they exist.

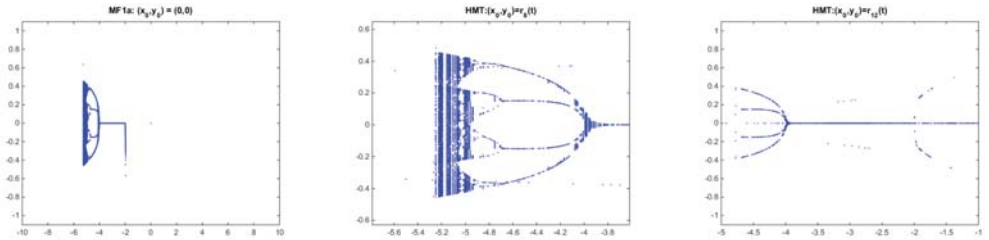


(a) $(0,0)$ (b) r_1 (c) r_2

Figure 3: Bifurcation diagrams of family MF on $p_1(x)$ starting from fixed points

To draw these Feigenbaum diagrams, 500 elements of the orbit of each strange fixed point are calculated, plotting the last 400, for each value of parameter θ (after a partition of the analyzed interval in 5000 subintervals). To avoid unnecessary repetitions, we do not show the bifurcation diagrams of all the strange fixed points. It has been checked that the orbits of these fixed points depend on the stability of the origin (some of them tend to $(0,0)$ for $\theta \in]-4,-2[$), but mostly tend to the roots of $p(x)$. However, some details of these bifurcation diagrams must be calculated, as it seems chaotic behavior is observed in some figures.

By using the strange fixed point $(0,0)$ as initial estimation, a bifurcation diagram can be seen in Figure 3a. Let us remark in this case that it fully coincides with the stability of this fixed point, stated in Proposition 2; that is, $(0,0)$ is an attracting fixed point for $\theta \in]-4,-2[$. Out of this interval iterations (points in blue color) tend to one of the roots, -1 or 1.



(a) $(0,0)$ (b) r_8 (c) r_{12}

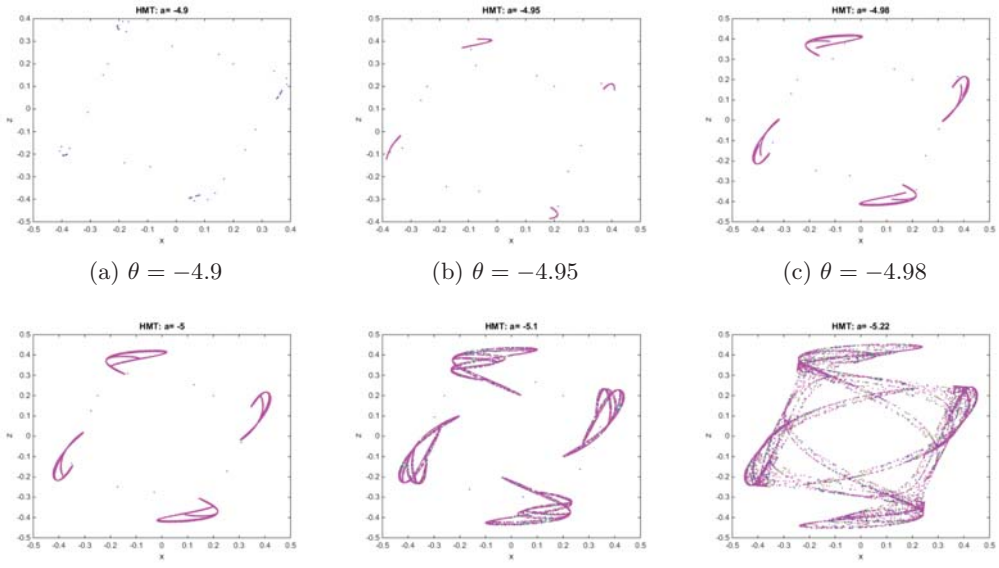
Figure 4: Bifurcation diagrams of family MF on $p_2(x)$ starting from fixed points

 ISBN: 978-84-608-6082-2

In Figure 3b, some unstable behavior appears in the same conflictive area (interval $\theta \in [-6,-4]$) where strange attractors appeared in the corresponding family of iterative methods without derivatives. Further analysis is necessary to determine if there exist also here this kind of behavior. Moreover, when other strange initial points are used as initial estimations, convergence yields to one of the roots of the polynomial $p_1(x)$, or they become complex.

When bifurcation diagrams are plotted in case of strange fixed points of operator $M_{2,\theta}(z,x)$, doubling-period bifurcation are found for values $\theta \in]-5,-4[$ in case of $(0,0)$, r_6, r_8 or r_{12} are used as initial estimations (see Figure 4).

To clarify the behavior of family MF with memory, we plot in (x,z)-space the iteration of operator $M_{2,\theta}(z,x)$, for values of parameter θ in specific blue regions of Figure 4b. So, some symmetric attractors have been found, (see Figure 5). The way these pictures have been obtained is the following: fixing the value of parameter θ, 10000 different initial estimations have been taken in rectangle $[-0.5,0.5]\times[-0.5,0.5]$. The method has been used on each of them, plotting one point per iteration. The resulting images show that four attracting curves appear and it can be observed that small perturbations in the value of the parameter make these attracting regions bigger until they become one only strange attractor.



(a) $\theta = -4.9$ (b) $\theta = -4.95$ (c) $\theta = -4.98$

(d) $\theta = -5$ (e) $\theta = -5.1$ (f) $\theta = -5.22$

Figure 5: Strange attractors of family MF on $p_2(x)$

 ISBN: 978-84-608-6082-2

Acknowledgements

This research was partially supported by Ministerio de Economía y Competitividad MTM2014-52016-C02-2-P.

References

[1] B. Campos, A. Cordero, J. R. Torregrosa and P. Vindel, *A multidimensional dynamical approach to iterative methods with memory*, Appl. Math. Comput. **271** (2015) 701-715.

[2] Z. Elhadj and J. C. Sprott, *On the Dynamics of a New Simple 2-d Rational Discrete Mapping*, Int. Bif. Chaos **21**(1) (2011) 1–6.

[3] W.F. Hassan Al-Shameri, *Dynamical properties of the Hénon mapping*, Int. Math. Anal. **6**(49) (2012) 2419–2430.

[4] J.L. Hueso, E. Martínez and J. R. Torregrosa, *New modifications of Potra-Pták's method with optimal fourth and eighth orders of convergence*, Comput. Appl. Math. **234** (2010) 2969–2976.

[5] V. G. Ivancevic, T. T. Ivancevic, *High-Dimensional Chaotic and Attractor Systems: A Comprehensive Introduction*, Springer Science & Business Media, 2007.

[6] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, 1970.

[7] R. C. Robinson, *An Introduction to Dynamical Systems, Continous and Discrete*, American Mathematical Society, Providence, 2012.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Approximating the matrix sign function by means of Chebyshev-Halley type method

Alicia Cordero[1], Fazlollah Soleymani[1], Juan R. Torregrosa[1] and M. Zaka Ullah[2]

[1] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, 46022 València, Spain*

[2] *Department of Mathematics, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

emails: acordero@mat.upv.es, fazlollah.soleymani@gmail.com., jrtorre@mat.upv.es, mzhussain@kau.edu.sa.

Abstract

This study presents several matrix iterative methods for finding the sign of a square complex matrix. These schemes are obtained by transforming a family of parametric iterative methods for solving nonlinear equations to the context of nonlinear matrix equations. Several special cases including global convergence behavior are dealt with. We show that they are asymptotically stable and have order of convergence six. Several numerical experiments for random matrices with different sizes are presented. These results show the effectiveness of the proposed schemes.

Key words: Matrix sign function, stability, iterative methods, Chebyshev-Halley family, eigenvalues.

1 Introduction

It is known that the function of sign in the scalar case is defined for any $z \in \mathbb{C}$ not on the imaginary axis by

$$sign(z) = \begin{cases} 1, & \mathrm{Re}(z) > 0, \\ -1, & \mathrm{Re}(z) < 0. \end{cases}$$

An extension of this function for the matrix case was given firstly by Roberts in [14], who introduced the matrix sign function as a tool for model reduction and for solving Lyapunov and algebraic Riccati equations. The problem of computing a function of a matrix, named

by $f(A)$, is of growing significance, though as yet numerical methods are developed for this purpose. In between, matrix sign function has undoubtedly a clear importance in the theory and application of matrix functions (e.g. one may refer to [3, 4, 15]). The matrix sign function has basic theoretical and algorithmic relations with the matrix square root, the polar decomposition, and from time to time, with the matrix pth roots (see for more [7, chapter 5]).

The matrix sign function is a valuable tool for the numerical solution of Sylvester and Liapunov matrix equations [1]. A generalization of the Newton iteration for the matrix sign function to the solution of the generalized algebraic Bernoulli equations was presented in [2]. This matrix function is used in [13] as a simple and direct method to derive some fundamental results in the theory of surface waves in anisotropic materials. For other applications of matrix sign function, we refer the reader to [12]. Due to the applicability of the matrix sign function, stable iterative schemes have become some viable choices for approximating this matrix.

Assume that $A \in \mathbb{C}^{n\times n}$ is a matrix with no eigenvalues on the imaginary axis. To define this matrix function formally, let $A = PJP^{-1}$ be a Jordan canonical form arranged so that $J = diag(J_1, J_2)$, where the eigenvalues of $J_1 \in \mathbb{C}^{p\times p}$ lie in the open left half-plane and those of $J_2 \in \mathbb{C}^{n-p\times n-p}$ lie in the open right-plane, then

$$S = sign(A) = P \begin{pmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{pmatrix} P^{-1}.$$

This matrix function can be uniquely defined (A is a nonsingular square matrix). The most concise definition of the matrix sign decomposition is given in [5, 10] as follows:

$$A = SN = A(A^2)^{-1/2}(A^2)^{1/2}, \tag{1}$$

whereas $S = A(A^2)^{-1/2}$ is the matrix sign function and $1/2$ denotes the principal matrix square root of a given matrix.

This matrix function has several properties. Some of them are given by [10]:

1. $S^2 = I$ (S is involutory).
2. S is diagonalizable with eigenvalues ± 1.
3. $SA = AS$.
4. If A is real, then S is real.
5. $(I + S)/2$ and $(I - S)/2$ are projectors onto the invariant subspaces associated with the eigenvalues in the right half-plane and left half-plane, respectively.

A number of matrix functions $f(A)$ are amenable to computation by iteration functions of the form $X_{k+1} = g(X_k)$ [7], where for the iterations used in practice, X_0 is not arbitrary but is a fixed function of A. Taking into account the computational burden makes it obvious that g is a polynomial or rational function. Rational g requires the solution of linear systems with multiple right-hand sides, or even explicit matrix inversion.

It is necessary to recall that outcomes and intuition from scalar nonlinear iterations do not necessarily generalize to the matrix case. As an illustration, standard convergence conditions expressed in terms of derivatives of g at a fixed point in the scalar case do not directly translate into analogous conditions on the Frechét and higher order derivatives in the matrix case.

In this paper, we focus on iterative methods for finding matrix S. In fact, such methods are Newton-type ones which are fixed-point methods producing a convergent sequence of matrices via applying a suitable initial matrix. The most known method of this class is the quadratic Newton's method (NM) defined by

$$X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-1}\right), \tag{2}$$

when $X_0 = A$ has been chosen as an initial matrix.

It should be remarked that iterative methods, such as (2) and Newton-Schultz iteration (NSM)

$$X_{k+1} = \frac{1}{2}X_k\left(3I - X_k^2\right), \tag{3}$$

or the cubically convergent Halley method (HM)

$$X_{k+1} = \left[I + 3X_k^2\right]\left[X_k(3I + X_k^2)\right]^{-1}, \tag{4}$$

are special cases of the Padé family or its reciprocal proposed originally in [11].

Motivated by the recent developments in this area [6, 16], we here propose some computationally variants of Chebyshev-Halley type scheme possessing a free parameter. An improvement of this family is given as our main contribution to possess high rate of convergence with global convergence behavior for some of its special members. The stability of the schemes are considered to show that the rounding errors remain under control.

The rest of the paper is organized as follows. In Section 2, a Chebyshev-Halley type family of schemes for solving nonlinear scalar equations is proposed. Some discussions on several members of the family are given, including the global convergence, the stability and the order of convergence. In Section 3, various numerical examples are considered to confirm the theoretical results. A comparison with the existing methods is also presented therein.

 ISBN: 978-84-608-6082-2

2 Construction of the family of schemes

It is known that Chebyshev-Halley family of iterative methods for approximating the simple roots of the nonlinear equation $f(x) = 0$ has the iterative expression

$$x_{k+1} = x_k - \left(1 + \frac{1}{2}\frac{L(x_k)}{1 - aL(x_k)}\right)\frac{f(x_k)}{f'(x_k)}, \tag{5}$$

where $a \in \mathbb{R}$ is an arbitrary parameter and $L(x_k) = \dfrac{f''(x_k)f(x_k)}{f'(x_k)^2}$ is the logarithmic convexity operator. For every value of parameter a, the convergence order is cubic. Here, we consider the general expression (5) to solve the following nonlinear matrix equation $X^2 = I$, where I is the identity matrix, and we obtain the following iterative expression in the reciprocal form

$$X_{k+1} = \left(-4aX_k + 4(-2+a)X_k^3\right)\left[I - 3X_k^2(2I + X_k^2) + 2a(-I + X_k^4)\right]^{-1}. \tag{6}$$

The main goal and motivation in constructing iterative methods for matrix sign is to attain as fast as possible order of convergence with minimal computational costs.

Expression (6) is costly to attain matrix sign function since it requires four matrix matrix multiplications (mmm) and one inverse per computing cycle to reach the local order 3. To improve this family of schemes and construct a more economic family of iterations, we propose the following nonlinear scalar solver,

$$\begin{aligned} y_k &= x_k - \left(1 + \frac{1}{2}\frac{L(x_k)}{1 - aL(x_k)}\right)\frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= y_k - \frac{f(y_k)}{f[y_k, x_k]}, \end{aligned} \tag{7}$$

where $f[y_k, x_k] = \dfrac{f(y_k) - f(x_k)}{y_k - x_k}$. By using Taylor expansion of the different elements of the iterative expression we can prove the following result.

Theorem 1 *Let $f(x)$ be at least three times differentiable in a neighborhood of its simple zero α. If an initial approximations x_0 is sufficiently close to α, then sequence $\{x_k\}_{k\geq 0}$ obtained from (7) converge to α with at least order of convergence four.*

From the reciprocal form of (7) we design a new family of improved Chebyshev-Halley type iterative methods for solving $X^2 - I = 0$:

$$X_{k+1} = \left(X_k - 6aX_k + 2(-7+2a)X_k^3 + (-3+2a)X_k^5\right)W_k^{-1}, \tag{8}$$

where $W_k = \left[(1-2a)I - 2(3+2a)X_k^2 + (-11+6a)X_k^4\right]$.

 ISBN: 978-84-608-6082-2

Further simplifying results to

$$X_{k+1} = X_k \left((1-6a)I + 2(-7+2a)X_k^2 + (-3+2a)X_k^4\right) W_k^{-1}, \tag{9}$$

which requires four mmm and one matrix inverse to reach a higher rate of convergence four in contrast to (6). Note that, X_k $(k \geq 0)$, are rational functions of A and hence, like A, commute with S.

On modern computers with hierarchical memories, matrix multiplication is usually much faster than solving a matrix equation or inverting a matrix, so iterations such as (9) that are multiplication-rich are preferred. In what follows, we list some special cases from the family (9).

- Choosing $a = 0$ results in (PM1):
  $X_{k+1} = X_k \left(-I + 14X_k^2 + 3X_k^4\right) \left[-I + 6X_k^2 + 11X_k^4\right]^{-1}$.
- Choosing $a = 1/2$ results in (PM2):
  $X_{k+1} = \left(I + 6X_k^2 + X_k^4\right) \left[4(X_k + X_k^3)\right]^{-1}$.
- Choosing $a = -1/2$ results in (PM3):
  $X_{k+1} = X_k \left(-2I + 8X_k^2 + 2X_k^4\right) \left[-I + 2X_k^2 + 7X_k^4\right]^{-1}$.
- Choosing $a = 1$ results in (PM4):
  $X_{k+1} = X_k \left(5I + 10X_k^2 + X_k^4\right) \left[I + 5X_k^2(2 + X_k^2)\right]^{-1}$.
- Choosing $a = -1$ results in (PM5):
  $X_{k+1} = X_k \left(-7I + 18X_k^2 + 5X_k^4\right) \left[-3I + 2X_k^2 + 17X_k^4\right]^{-1}$.
- Choosing $a = -2$ results in (PM6):
  $X_{k+1} = X_k \left(-13I + 22X_k^2 + 7X_k^4\right) \left[-5I - 2X_k^2 + 23X_k^4\right]^{-1}$.
- Choosing $a = 3/2$ results in (PM7):
  $X_{k+1} = X_k \left(4(I + X_k^2)\right) \left[I + 6X_k^2 + X_k^4\right]^{-1}$.
- Choosing $a = -3/2$ results in (PM8):
  $X_{k+1} = X_k \left(-5I + 10X_k^2 + 3X_k^4\right) \left[-2I + 10X_k^4\right]^{-1}$.
- Choosing $a = -4/5$ results in (PM9):
  $X_{k+1} = X_k \left(-29I + 86X_k^2 + 23X_k^4\right) \left[-13I + 14X_k^2 + 79X_k^4\right]^{-1}$.

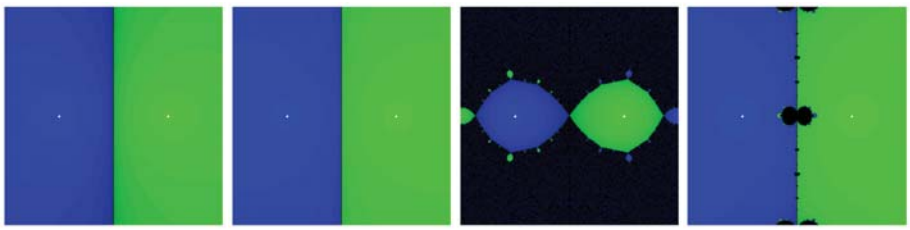
 ISBN: 978-84-608-6082-2

It is quite obvious that the sign matrix may be used to determine the number of eigenvalues of a given matrix A to the right or left of any straight line $x = a$, $(a \in \mathbb{R})$ in the complex (x, y) plane [8]. To be more precise, the above iterations may be used to determine whether a matrix is stable. It is also apparent that we may easily determine the number of eigenvalues inside a vertical strip bounded by the lines $x = b$ and $x = c$ with $b, c \in \mathbb{R}$ and $b < c$, provided that no eigenvalues of A lie on these lines.

In the rest of this section it is discussed for which values of the free parameter a, one may attain an efficient scheme for computing matrix sign function. We remark that a method for computing S must be globally convergent and it is of practical interest if it does not belong to the general Padé family.

So, it must be checked that for which values of a the convergence is global. To pursue this aim, it is enough to draw the basins of attraction for the scheme (9) to solve the scalar equation $f(x) := x^2 - 1 = 0$ (for more information on pure matrix methods and their global convergence behavior one should consult the thesis [9]). We take a rectangle $\mathbb{D} = [-2, 2] \times [-2, 2] \subset \mathbb{C}$ and assign a color to each point $z_0 \in \mathbb{D}$ according to the simple zero at which the scheme from (9) converges and we mark the point as black if the method does not converge. Here, we take into account the stopping criterion for convergence to be $|f(x_k)| \leq 10^{-2}$ wherein the maximum number of full cycles for each method is 200 in the written Mathematica codes. Following such a procedure, we distinguish the attraction basins by their colors for different methods, which are shaded according to the number of iterations.

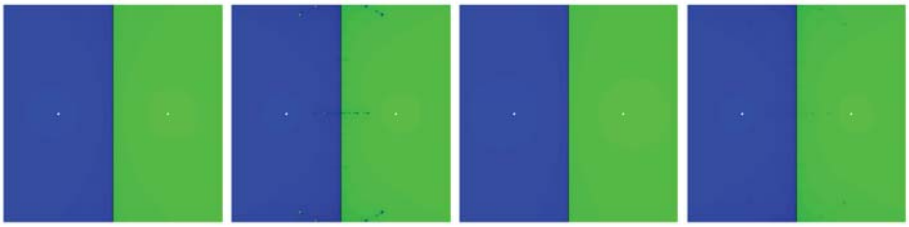
Results of dynamical behaviors for different cases are brought forward in Figures 1-3. Checking the results and comparing by the schemes from Padé family, it is yield that PM1, PM3, PM5, PM8 and PM9 are not of global convergence and they would not be of further interest. On the other hand, PM2, PM4 and PM7 are members from the Padé family. While this shows the generality of the proposed Chebyshev-Halley type method, we can deduce that PM6 is new with global convergence.



(a) NM (b) HM (c) NSM (d) $PM1$

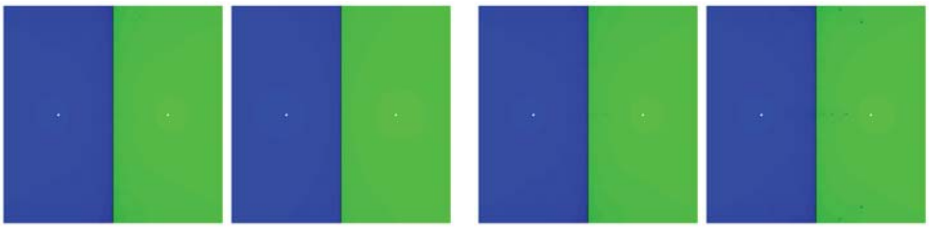
Figure 1: Basins of atraction for Newton, Halley, Newton-Schulz methods and $a = 0$

 ISBN: 978-84-608-6082-2



(a) $PM2$ (b) $PM3$ (c) $PM4$ (d) $PM5$

Figure 2: Basins of atraction for $a = 0.5$, $a = -0.5$, $a = 1$ and $a = -1$



(a) $PM6$ (b) $PM7$ (c) $PM8$ (d) $PM9$

Figure 3: Basins of atraction for $a = -2$, $a = 3/2$, $a = -3/2$ and $a = -4/5$

It can be proved that several members of the proposed family of Chebyshev-Halley type schemes (9) are convergent with fourth-order, by choosing as initial guess $X_0 = A$. In addition, it can be analyzed how a small perturbation at kth iterate is amplified or damped along the iterates, which could be considered as asymptotical stability.

3 Numerical experiments

In this section we present some numerical tests for computing the matrix sign function, by using Mathematica 8 with 53 digit binary. In this work, the computer specifications are Windows 7 Ultimate with Intel(R) Core(TM) i5-2430M CPU 2.40GHz processor and 8.00 GB of RAM on a 64-bit operating system.

Different methods are compared in terms of number of iterations and the computational CPU time. We only apply methods with global convergence behavior for comparison. The compared schemes are NM, HM, PM4, PM7, PM6 and ANM (accelerated Newton's method)

which is defined by

$$\begin{cases} X_0 = A, \\ \mu_k = \sqrt{\dfrac{\|X_k^{-1}\|}{\|X_k\|}}, \\ X_{k+1} = \frac{1}{2}\left(\mu_k X_k + \mu_k^{-1} X_k^{-1}\right). \end{cases} \tag{10}$$

One can similarly accelerate the performance of the new schemes from the improved Chebyshev-Halley type family (9) using some strategy as in (10). But since the computation of the scaling parameter μ_k is occasionally costly, we do not study it deeply for our family of iterations. The stopping criterium in this work is $\|X_k^2 - I\|_2 \leq 10^{-8}$.

Example 1 *In this series of experiments, we compute the matrix sign function of the following 10 randomly generated matrices*

```
SeedRandom[1234]; number = 10;
Table[A[l] = RandomReal[{-100, 100}, {100 l, 100 l}];, {l, number}];
```

The results are displayed in Tables 1-2 on random matrices of size $100i \times 100i$, $i = 1, 2, \ldots, 10$. The results confirm the theoretical aspects of Sections 2-3. They show that there is a reduction in the number of iterations and computational time using PM4, PM7 and PM6. PM4 and PM7 are the best methods in terms of computational times. Note that the computation of X_k^2 per cycle for calculating the stopping condition adds one matrix-matrix multiplication for NM, while the HM and the proposed methods form this matrix during the process of each step.

Matrix No.	NM	ANM	HM	PM7	PM4	PM6
$A_{100\times100}$	17	11	11	9	8	10
$A_{200\times200}$	19	14	12	10	8	12
$A_{300\times300}$	20	16	13	10	9	12
$A_{400\times400}$	24	18	15	12	11	14
$A_{500\times500}$	20	16	13	10	9	12
$A_{600\times600}$	23	21	14	12	10	14
$A_{700\times700}$	22	18	14	11	10	13
$A_{800\times800}$	23	21	15	12	10	14
$A_{900\times900}$	23	19	14	12	10	14
$A_{1000\times1000}$	23	21	15	12	10	14

Table 1: Comparison of number of iterations for Example 1.

Matrix No.	NM	ANM	HM	PM7	PM4	PM6
$A_{100\times100}$	0.0368623	0.0531857	0.0275516	0.0263713	0.0235674	0.0269054
$A_{200\times200}$	0.158083	0.273685	0.121207	0.113714	0.0977962	0.135305
$A_{300\times300}$	0.44521	0.843781	0.353374	0.315238	0.289896	0.367553
$A_{400\times400}$	1.09422	1.91654	0.88232	0.842601	0.829421	0.984137
$A_{500\times500}$	1.57968	2.92306	1.37999	1.19248	1.1876	1.45492
$A_{600\times600}$	2.97305	6.32247	2.37401	2.30263	2.16222	2.7574
$A_{700\times700}$	4.54777	8.51585	3.7262	3.3351	3.27217	3.94201
$A_{800\times800}$	7.25574	15.3073	6.00425	5.25829	4.87302	6.25886
$A_{900\times900}$	10.361	20.1839	8.11459	7.57771	6.7404	9.08985
$A_{1000\times1000}$	14.3216	31.2905	11.6792	10.3742	9.34322	12.3002

Table 2: Comparison of the elapsed time for Example 1.

Similar numerical experiments have been carried out on variety of problems which confirm the above conclusions to a great extent. Finally, we can conclude from numerical experiments that new proposed schemes confirm the theoretical results and show consistent convergence behavior.

Acknowledgements

This research was partially supported by Ministerio de Economía y Competitividad MTM2014-52016-C02-2-P.

References

[1] P. Benner, E.S. Qintana-Ortí, *Solving stable generalized Lyapunov equations with the matrix sign function*, Numer. Algor. **20**(1) (1999) 75–100.

[2] S. Barrachina, P. Benner, E.S. Qintana-Ortí, *Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function* , Numer. Algor. **46**(4) (2007) 351–368.

[3] F.D. Filbir, *Computation of the structured stability radius via matrix sign function*, Sys. & Contr. Lett. **22** (1994) 341–349.

[4] O. Gomilko, F. Greco, K. Zięţak, *A Padé family of iterations for the matrix sign function and related problems*, Numer. Lin. Alg. Appl. **19** (2012) 585–605.

[5] N.J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Alg. Appl. **212/213** (1994) 3–20.

[6] N.J. Higham, D.S. Mackey, N. Mackey, F. Tisseur, *Computing the polar decomposition and the matrix sign decomposition in matrix groups*, SIAM J. Matrix Anal. Appl. **25** (2004) 1178–1192.

[7] N.J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[8] J.L. Howland, *The sign matrix and the separation of matrix eigenvalues*, Linear Alg. Appl. **49** (1983) 221–232.

[9] B. Iannazzo, Numerical solution of certain nonlinear matrix equations, Ph.D. thesis, Dipartimento di Matematica, Università di Pisa, 2007.

[10] C. Kenney, A.J. Laub, em Rational iterative methods for the matrix sign function, SIAM J. Matrix Anal. Appl. **12** (1991) 273–291.

[11] C.S. Kenney, A.J. Laub, *The matrix sign function*, IEEE Trans. Automat. Cont. **40** (1995) 1330–1348.

[12] M.Sh. Misrikhanov, V.N. Ryabchenko, *Matrix sign function in the problems of analysis and design of the linear systems*, Auto. Remote Control **69** (2008) 198–222.

[13] A.N. Norris, A.L. Shuvalov, A.A. Kutsenko, *The matrix sign function for solving surface wave problems in homogeneous and laterally periodic elastic half-spaces*, Wave Motion **50**(8) (2013) 1239–1250.

[14] J.D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function* Int. J. Control. **32** (1980) 677–687.

[15] A.R. Soheili, F. Toutounian, F. Soleymani, *A fast convergent numerical method for matrix sign function with application in SDEs*, Comput. Appl. Math. **282** (2015) 167–178.

[16] F. Soleymani, P.S. Stanimirović, I. Stojanović, *A novel iterative method for polar decomposition and matrix sign function*, Disc. Dyn. Nature Soc. **2015** Art. ID 649423, 11 pages.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Dynamical study of Ostrowski' and Chun's methods for solving nonlinear systems

Alicia Cordero[1], Javier G. Maimó[2], Juan R. Torregrosa[1] and María P. Vassileva[2]

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

[2] *Instituto Tecnológico de Santo Domingo (INTEC), Santo Domingo, República Dominicana*

emails: acordero@mat.upv.es, javier.garcia@intec.edu.do, jrtorre@mat.upv.es, maria.penkova@intec.edu.do

Abstract

In this paper we present a comparison between the real multidimensional dynamical behaviour of Ostrowski' and Chun's methods to solve systems of nonlinear equations. In general, scalar methods can be transfered to make them suitable to solve nonlinear systems. The dynamical behavior of the rational operator associated to a scalar method applied to low-degree polynomials has shown to be an efficient tool for analyzing the stability and reliability of the methods. However, although it is possible to transfer both Ostrowski' and Chun's methods from equations to systems of equations, a good scalar dynamical behaviour does not guarantee a good one in multidimensional case.

Key words: Nonlinear system of equations, iterative method, basin of attraction, dynamical plane, stability.

1 Introduction

Let us consider the problem of finding a real zero of a function $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ that is, a solution $\bar{x} \in D$ of the nonlinear system $F(x) = 0$, of n equations with n variables, being f_i, $i = 1, 2, \ldots, n$ the coordinate functions of F. This solution can be obtained as a fixed point of some function $\bar{G} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by means of the fixed-point iteration method

$$x^{(k+1)} = \bar{G}(x^{(k)}), \quad k = 0, 1, \ldots, \tag{1}$$

 ISBN: 978-84-608-6082-2

where $x^{(0)}$ is the initial estimation.

Methods for solving nonlinear equations $f(x) = 0$, $f : I \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ can be transferred to systems $F(x) = 0$, $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$. To adapt a scalar method a variable is replaced by a vector, but not all methods are easy to adapt, the extension to systems requires to rewrite the iterative expression in such a way that there are no evaluations of the nonlinear function F in the denominator, as they will become vectors in the extension to systems. To solve this problem the divided difference operator can be used $[x, y; F]$, see [1]. Once the method has been transfered to multivariate case, a dynamical study can be made to see if good methods for solving nonlinear equations are still stable when extended to systems.

In [1] a new family is introduced, the Ostrowski-Chun family of iterative methods, and the class is extended to systems by using que divided difference operator. In this paper we analyze the real multidimensional dynamical behaviour of two members of this family.

This paper is organized as follows: in Section 2 some concepts of real multidimensional dynamics extended to nonlinear systems are introduced. In Section 3 we study the dynamical behaviour of Ostrowski' and Chun's methods. Finally, some conclusions are stated.

2 Basic concepts

Some dynamical studies by using complex dynamics tools have been made for iterative methods for solving nonlinear equations on low degree polynomials, see [2, 3, 4]. These techniques have proved to be efficient to analyze the stability of a method or to select the most stable members of a family.

In this work, we propose a dynamical study of Ostrowski' and Chun's method to solve systems of nonlinear equations. In order to analyze the dynamical behavior of a fixed-point iterative method for nonlinear systems when is applied to n-variable polynomial $p(x)$, $p : \mathbb{R}^n \to \mathbb{R}^n$, $x \in \mathbb{R}^n$, it is necessary to recall some basic dynamical concepts (see for example [5, 6]).

Let us denote by $G(x)$ the vectorial fixed-point function associated to the iterative method on polynomial $p(x)$. Let us note that the next concepts and results are also valid when the iterative method is applied on a general function $F(x)$.

Definition 2.1 *Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be a vectorial rational function. The orbit of a point $x^{(0)} \in \mathbb{R}^n$ is defined as the set of successive images of $x^{(0)}$ by the vectorial rational function, $\left\{x^{(0)}, G(x^{(0)}), \ldots, G^m(x^{(0)}), \ldots\right\}$.*

The dynamical behavior of the orbit of a point of $\mathbb{R}^n$ can be classified depending on its asymptotic behavior. In this way, a point $x^* \in \mathbb{R}^n$ is a fixed point of G if $G(x^*) = x^*$.

We recall a known result in discrete dynamics that gives the stability of fixed or periodic points for nonlinear operators.

Theorem 1 ([6], page 558) *Let G from $\mathbb{R}^n$ to $\mathbb{R}^n$ be $\mathcal{C}^2$. Assume x^* is a period-k point. Let $\lambda_1, \lambda_1, \ldots, \lambda_n$ be the eigenvalues of $G'(x^*)$.*

a) If all the eigenvalues λ_j have $|\lambda_j| < 1$, then x^* is attracting.

b) If one eigenvalue λ_{j_0} has $|\lambda_{j_0}| > 1$, then x^* is unstable, that is, repelling or saddle.

c) If all the eigenvalues λ_j have $|\lambda_j| > 1$, then x^* is repelling.

In addition, a fixed point is called hyperbolic if all the eigenvalues λ_j of $G'(x^*)$ have $|\lambda_j| \neq 1$. In particular, if there exist an eigenvalue λ_i such that $|\lambda_i| < 1$ and an eigenvalue λ_j such that $|\lambda_j| > 1$, the hyperbolic point is called saddle point.

Let us note that, the entries of $G'(x^*)$ are the partial derivatives of each coordinate function of the vectorial rational operator that defines the iterative scheme.

If a fixed point is not a root of the nonlinear function, it is called strange fixed point and its character can be analyzed in the same manner. Then, if x^* is an attracting fixed point of the rational function G, its basin of attraction $\mathcal{A}(x^*)$ is defined as the set of pre-images of any order such that

$$\mathcal{A}(x^*) = \left\{ x^{(0)} \in \mathbb{R}^n : G^m(x^{(0)}) \to x^*, m \to \infty \right\}.$$

As in the scalar case, the set of points whose orbits tend to an attracting fixed point x^* is defined as the Fatou set, $\mathcal{F}(G)$. The complementary set, the Julia set $\mathcal{J}(G)$, is the closure of the set consisting of its repelling fixed points, and establishes the borders between the basins of attraction.

3 Dynamical study of Ostrowski' and Chun's methods

In this section we will apply the previous dynamical concepts to the rational functions associated to Ostrowski' and Chun's methods, two well known 4th-order methods included in the family of biparametric schemes designed in [1].

Ostroswski's method is an optimal fourth order method, when applied to scalar polynomials of degree 2 the basins of attraction are connected and its Fatou set is the same as Newton's one but with fourth order of convergence. Ostrowski's method transferred to systems takes the form:

$$x^{(k+1)} = y^{(k)} - \left(-I + 2[F'(x^{(k)})]^{-1}[x^{(k)}, y^{(k)}; F]\right)^{-1} [F'(x^{(k)})]^{-1} F(y^{(k)}). \tag{2}$$

where $y^{(k)}$ is Newton's step.

 ISBN: 978-84-608-6082-2

On the other hand Chun's method is an optimal method of order 4, when applied to scalar polynomials of degree 2 shows no instability. The iterative expression of Chun's method transferred to systems is:

$$x^{(k+1)} = y^{(k)} - \left(I - 2[F'(x^{(k)})]^{-1}[x^{(k)}, y^{(k)}; F]\right)[F'(x^{(k)})]^{-1}F(y^{(k)}). \tag{3}$$

where $y^{(k)}$ is Newton's step.

In particular, we will analyze the dynamical behavior of the methods acting on the polynomial systems $p(x) = 0$ and $q(x) = 0$, where

$$\left.\begin{array}{l} p_1(x) = x_1^2 - 1 \\ p_2(x) = x_2^2 - 1 \end{array}\right\}, \qquad \left.\begin{array}{l} q_1(x) = x_1x_2 + x_1 - x_2 - 1 \\ q_2(x) = x_1x_2 - x_1 + x_2 - 1 \end{array}\right\}.$$

3.1 Analysis of the fixed points

Now we will substitute our test systems $p(x)$ and $q(x)$ in the iterative expressions of Ostrowski' and Chun's methods, (2) and (3), to study its fixed points.

3.1.1 Fixed point operators on $p(x)$

The jth-coordinate of the vectorial rational function associated to Ostrowski' scheme on polynomial $p(x)$ is

$$\bar{\lambda}_j^{Os,p}(x) = \frac{1}{4}\frac{x_j^4 + 6x_j^2 + 1}{x_j(x_j^2 + 1)} \quad j = 1, 2. \tag{4}$$

The fixed points are the solutions of $\bar{\lambda}_j^{Os,p}(x) = x_j$

$$\frac{1}{4}\frac{x_j^4 + 6x_j^2 + 1}{x_j(x_j^2 + 1)} = x_j \ , \ j = 1, 2,$$

that is

$$\frac{1}{4}(x_j^2 - 1)(3x_j^2 + 1) = 0 \ , \ j = 1, 2.$$

Let us remark that has only two real solutions: -1,1. As we have two possible values for j that leads to 4 fixed points, the roots of $p(x)$, $(1, 1)$, $(1, -1)$, $(-1, 1)$ and $(-1, -1)$.

To check their stability we need to calculate the Jacobian matrix and evaluate their eigenvalues in each fixed point,

$$J^{Os,p}(x_1, x_2) = \begin{pmatrix} \dfrac{\left({x_1}^2 - 1\right)^3}{4{x_1}^2\left({x_1}^2 + 1\right)^2} & 0 \\ 0 & \dfrac{\left({x_2}^2 - 1\right)^3}{4{x_2}^2\left({x_2}^2 + 1\right)^2} \end{pmatrix}$$

 ISBN: 978-84-608-6082-2

with eigenvalues

$$\left(\frac{\left({x_1}^2-1\right)^3}{4{x_1}^2\left({x_1}^2+1\right)^2},\frac{\left({x_2}^2-1\right)^3}{4{x_2}^2\left({x_2}^2+1\right)^2}\right)^T.$$

It is clear that by evaluating these eigenvalues in the roots of the polynomial we obtain (0,0), which means they are attractive.

In an analogous way, the jth-coordinate of the vectorial rational function to Chun's scheme on polynomial $p(x)$ is

$$\bar{\lambda}_j^{Ch,p}(x)=\frac{1}{16}\frac{5x_j^6+15x_j^4-5x_j^2+1}{x_j^5}\quad j=1,2. \tag{5}$$

If we do the same analysis we did for Ostrowski's scheme the only real fixed points for are the roots of the $p(x)$, $(1,1)$, $(1,-1)$, $(-1,1)$ and $(-1,-1)$ with null eigenvalues, so they are attractive.

3.1.2 Fixed point operators on $q(x)$

For $q(x)$ we have for Ostrowski' scheme

$$\lambda_j^{Os,q}(x)=\frac{1}{2}\frac{x_1^2x_2^2+x_1^2+4x_1x_2+x_2^2+1}{x_1^2x_2+x_1x_2^2+x_1+x_2}\quad j=1,2.$$

fixed points are the solutions of

$$\frac{1}{2}\frac{x_1^2x_2^2+x_1^2+4x_1x_2+x_2^2+1}{x_1^2x_2+x_1x_2^2+x_1+x_2}=x_j\quad j=1,2.$$

The only real solutions of the system are (1,1) and (-1,-1), the roots of $q(x)$. By making the same analysis we did for $p(x)$ it can be seen that their eigenvalues are null, so they are attractive.

The coordinates of the vectorial rational function to for Chun's scheme on $q(x)$ are

$$\begin{aligned}\lambda_j^{Ch,q}(x) &= \frac{x_1^5x_2+2x_1^4x_2^2+3x_1^4+4x_1^3x_2^3+6x_1^3x_2+2x_1^2x_2^4+12x_1^2x_2^2}{(x_1+x_2)^5}\\ &+ \frac{-4x_1^2+x_1x_2^5+6x_1x_2^3-2x_1x_2+3x_2^4-4x_2^2+2}{(x_1+x_2)^5}\quad j=1,2.\end{aligned}$$

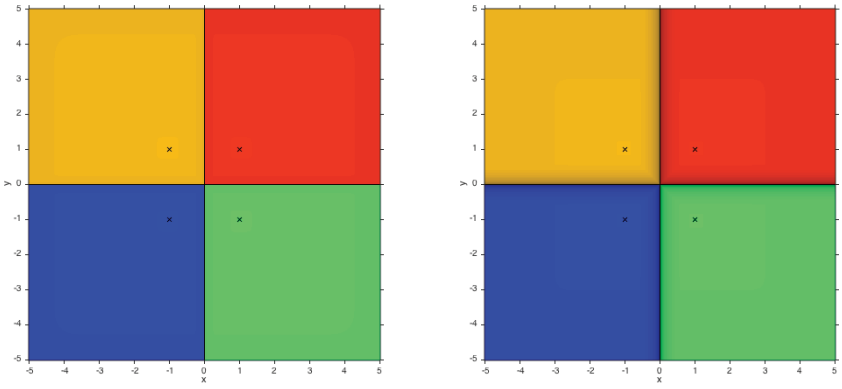
Again, the only real fixed points for are the roots of $q(x)$, $(1,1)$ and $(-1,-1)$, and they are attractive.

As there are no strange fixed points, we expect a good behaviour of the dynamical planes, as we see in the next section.

3.2 Dynamical planes

A dynamical plane is a visual representation of a method that gives qualitative information about its behaviour. Dynamical planes are built by applying the iterative method with different initial estimations distributed in a mesh. If the numerical method converges to a root of the system the point of the plane is painted in different colors depending on the root they converge to, while the points painted in black mean no convergence to any root after 40 iterations.

These dynamical planes have been obtained by using 400×400 subintervals, a maximum of 40 iterations and an error estimation of 10^{-3}, when the iterates tend to a fixed point.



(a) Ostrowski (b) Chun

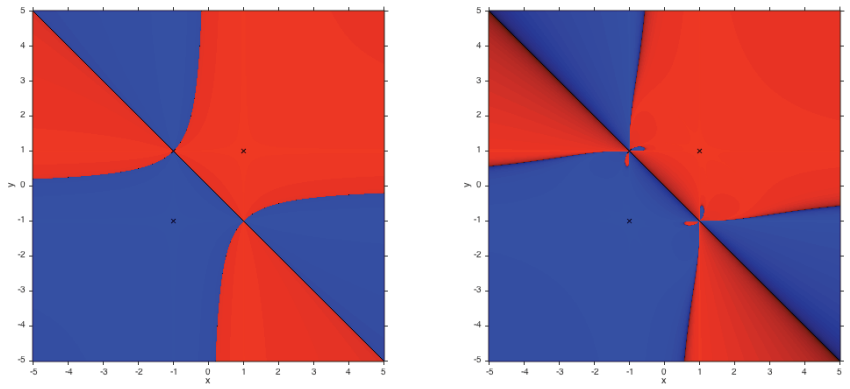
Figure 1: Dynamical plane for Ostroswki' and Chun's methods for $p(x)$

Figure 1 shows the dynamical plane for both schemes on $p(x)$. We can see four basins of attraction, one for each root of the polynomial. Both methods show a stable behaviour, as the only fixed points with a basin of attraction are the roots of the polynomial.

The dynamical planes of Ostroswki' and Chun's methods on $q(x)$ are shown in Figure 2. Again, both methods show stable behaviour and only the roots of the polynomial have basin of attraction.

4 Conclusions

Both Ostrowski' and Chun's methods show a good dynamical behavior when transferred to systems of equations. A complete study of all the Ostrowski-Chun's family of methods to select the best members of the family is necessary, in order to check if there exist other stable methods.



(a) Ostrowski (b) Chun

Figure 2: Dynamical plane for Ostroswki' and Chun's methods for $q(x)$

Acknowledgements

This research was supported by Ministerio de Economia y Competitividad MTM2014-52016-C02-02 and FONDOCYT 2014-1C1-088 Republica Dominicana.

References

[1] A. Cordero, J. García-Maimó, J. R. Torregrosa and M. P. Vassileva *Solving nonlinear problems by Ostrowski-Chun type parametric families*. Journal of Mathematical Chemistry, 53 (2015) 430-449.

[2] Á. A. Magreñán *Different anomalies in a Jarratt family of iterative root-finding methods*. Applied Mathematics and Computation, 233 (2014) 29–38.

[3] B. Neta, C. Chun, M. Scott *Basins of attraction for optimal eighth order methods to find simple roots of nonlinear equations* Applied Mathematics and Computation, 227 (2014) 567–592.

[4] A. Cordero, J. García-Maimó, J. R. Torregrosa, M. P. Vassileva and P. Vindel *Chaos in King's iterative family* Applied Mathematics Letters, 26(8) (2013) 842-848.

[5] A. Cordero and J. R. Torregrosa *Dynamical analysis of iterative methods for nonlinear systems or how to deal with the dimension?*. Applied Mathematics and Computation, 244 (2014) 398-412.

[6] R.C. Robinson *An Introduction to Dynamical Systems, Continous and Discrete* Americal Mathematical Society, Providence, 2012.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Computing the validity of attribute implications in multi-adjoint concept lattices

M. Eugenia Cornejo[1], Jesús Medina[2] and Eloísa Ramírez-Poussa[2]

[1] *Department of Statistic and O.R., University of Cádiz, Spain*

[2] *Department of Mathematics, University of Cádiz, Spain*

emails: mariaeugenia.cornejo@uca.es, jesus.medina@uca.es, eloisa.ramirez@uca.es

Abstract

Although attribute implications have widely been studied, the multi-adjoint framework offers powerful tools in order to improve the extracting attributes implications from a knowledge system. This paper extends a recent contribution developed on this framework, introducing an efficient mechanism in order to compute the validity of fuzzy implications in the multi-adjoint frame.

Key words: Concept lattice, attribute implication, join-irreducible element.

1 Introduction

Different research areas such as stock market prediction, disease diagnosis, census data analysis, etc., have a special interest in the extraction information from databases. Usually, a set of rules is obtained in order to summarize the obtained information from the knowledge of databases. These sets of rules are known as logic programs. However, traditional logic program languages are not able to handle uncertainty and approximated reasoning. In order to solve this trouble, different fuzzy logic programming systems have been proposed such as Probabilistic Deductive Databases, Hybrid Probabilistic Logic Programs, Annotated Logic Programming, Residuated Logic Programming, etc.

A generalization of previous logic programming frameworks was introduced, which is called Multi-adjoint Logic Programming [14, 15] and whose semantic structure is the multi-adjoint lattice. From a multi-adjoint lattice, a set of weighted rules and facts of a given language is defined, which is called multi-adjoint logic program. This work is focused on the use of a formal tool which allow us to obtain multi-adjoint logic programs from databases. Specifically, we are interested in employing Formal Concept Analysis.

 ISBN: 978-84-608-6082-2

Formal Concept Analysis is a mathematical technique to extract information from relational databases that contain a set of objects and a set of attributes. The pieces of the extracted information are called concepts and they are hierarchized to obtain concept lattices.

Attribute implications is an important topic in formal concept analysis in the classical case [3, 16] and in the fuzzy one [2, 8, 10, 11, 17]. Nevertheless, powerful advantages provided by multi-adjoint environment are not considered.

Our contribution complements the study presented in [12], related to attribute implications in multi-adjoint concept lattices [13], where the algebraic structure considered is a multi-adjoint algebra [6, 7] instead of the residuated lattice.

The paper is organized as follows. Section 2 shows the definitions and some properties of the calculation operators used in this paper. Section 3 establishes the main notions of the multi-adjoint concept lattices framework. The main results of this paper together an illustrative example are presented in Section 4. We prove that the degree in which a fuzzy implication is valid in a multi-adjoint concept lattice can be calculated by using of the set of join-irreducible elements of the lattice, instead of considering the whole set of intensions of the concepts of the lattice. Finally, the paper finishes with some conclusions and prospects for future work.

2 Adjoint triples and hedges

To begin with, we will introduce the operators that we will use to make the computations and some properties of them. First of all, we will present the notion of adjoint triple which arose as a generalization of a t-norm and its residuated implication [9]. The well-known adjoint property between a t-norm and its residuated implication is generalized in two different ways by using of adjoint triples.

Definition 1. *Let* $(P_1,\leq_1)$, $(P_2,\leq_2)$, $(P_3,\leq_3)$ *be posets and* $\&\colon P_1\times P_2\to P_3$, $\swarrow\colon P_3\times P_2\to P_1$, $\nwarrow\colon P_3\times P_1\to P_2$ *be mappings, then* $(\&,\swarrow,\nwarrow)$ *is an* adjoint triple *with respect to* P_1,P_2,P_3 *if the following equivalence:*

$$x\leq_1 z\swarrow y \quad \text{if and only if} \quad x\,\&\,y\leq_3 z \quad \text{if and only if} \quad y\leq_2 z\nwarrow x$$

holds, for all $x\in P_1$, $y\in P_2$ *and* $z\in P_3$. *The previous equivalence is called* adjoint property.

An interesting study of adjoint triples and their useful properties was included in [4]. Specifically, the proposition below states the conditions that the posets and the adjoint conjunctor must satisfy in order to ensure that the adjoint implications verify a property which can be seen as a generalization of the boundary conditions of the classical boolean implications.

Proposition 1 ([4])**.** *Given an adjoint triple* $(\&,\swarrow,\nwarrow)$ *with respect to the posets* $(P_1,\leq_1)$, $(P_2,\leq_2)$ *and* $(P_3,\leq_3)$.

(1) If $P_2 \subseteq P_3$ and P_1 has a maximum $\top_1$ as a left identity element for $\&$, that is, the equality $\top_1 \& y = y$ holds, for all $y \in P_2$, then we obtain

$$\top_1 = z \swarrow y \quad \text{if and only if} \quad y \leq_3 z$$

for all $y \in P_2$ and $z \in P_3$.

(2) If $P_1 \subseteq P_3$ and P_2 has a maximum $\top_2$ as a left identity element for $\&$, that is, the equality $x \& \top_2 = x$ is satisfied, for all $x \in P_1$, then we have

$$\top_2 = z \nwarrow x \quad \text{if and only if} \quad x \leq_3 z$$

for all $x \in P_1$ and $z \in P_3$.

Truth-stressing hedges were studied in [9] and they have also considered in the definition of the concept-forming operators in different fuzzy extensions of formal concept analysis [1].

Definition 2. *Let $(P_1, \leq_1, \top_1)$ be an upper bounded poset and $(\&, \swarrow, \nwarrow)$ be an adjoint triple with respect to P_1. A* truth-stressing a-hedge *(*a-hedge*, for short) is a unary function $*\colon P_1 \to P_1$ satisfying the properties: $\top_1^* = \top_1$, $a^* \leq_1 a$, $(b \swarrow a)^* = b^* \swarrow a^*$, $a^{**} = a$, for each $a, b \in P_1$.*

Similarly, a truth-stressing b-hedge *(*b-hedge*, for short) is defined on $(P_2, \leq_2, \top_2)$, only switching the property $(b \nwarrow a)^* = b^* \nwarrow a^*$.*

If the properties of both *a-hedges* and *b-hedges* are satisfied, we obtain the classic concept of hedge [9]. Two interesting hedges can be defined if a lattice $(L, \preceq)$ is considered. The identity hedge is defined as $a^* = a$, for all $a \in L$; and the globalization hedge, which is defined as:

$$a^* = \begin{cases} \top & a = \top \\ \bot & otherwise \end{cases}$$

for all $a \in L$. These hedges are the greater (identity) and the smaller (globalization) hedges.

The following result shows that the monotonicity of truth-stressing hedges is obtained if the adjoint conjunctor satisfies the boundary conditions.

Proposition 2. *Given an adjoint triple $(\&, \swarrow, \nwarrow)$ with respect to an upper bounded poset $(P, \leq, \top)$.*

(1) If the equality $x \& \top = x$ holds, for all $x \in P$, and the mapping $*\colon P \to P$ is a truth-stressing a-hedge, then $*$ is order-preserving.

(2) If the equality $\top \& y = y$ holds, for all $y \in P$, and the mapping $*\colon P \to P$ is a truth-stressing b-hedge, then $*$ is order-preserving.

 ISBN: 978-84-608-6082-2

3 Multi-adjoint concept lattices

The multi-adjoint concept lattice framework provides more flexibility into the language since different adjoint triples can be considered. Moreover, in this framework, operators do not need to be neither monotone nor associative. With the purpose of taking all advantages supplied by this environment, from now on, we will work with a multi-adjoint frame and a multi-adjoint context which are defined as follows.

Definition 3. *A multi-adjoint frame is a tuple* $(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$ *where* $(L_1, \preceq_1)$ *and* $(L_2, \preceq_2)$ *are complete lattices,* $(P, \leq)$ *is a poset and* $(\&_i, \swarrow^i, \nwarrow_i)$ *is an adjoint triple with respect to* L_1, L_2, P*, for all* $i \in \{1, \ldots, n\}$.

Definition 4. *Let* $(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$ *be a multi-adjoint frame, a context is a tuple* (A, B, R, σ) *such that A and B are non-empty sets (interpreted as attributes and objects, respectively), R is a P-fuzzy relation* $R\colon A \times B \to P$ *and* $\sigma\colon A \times B \to \{1, \ldots, n\}$ *is a mapping which associates any element in* $A \times B$ *with a particular adjoint triple in the frame.*

We will use the notation L_2^B and L_1^A to denote the set of fuzzy subsets $g\colon B \to L_2$ and $f\colon A \to L_1$ respectively. On these sets a pointwise partial order can be considered from the partial orders in $(L_1, \preceq_1)$ and $(L_2, \preceq_2)$, which provides L_2^B and L_1^A the structure of complete lattice. Given a multi-adjoint frame and a context, the concept-forming operators $^{\uparrow}\colon L_2^B \to L_1^A$ and $^{\downarrow}\colon L_1^A \to L_2^B$ are defined as:

$$\begin{aligned} g^{\uparrow}(a) &= \inf\{R(a,b) \swarrow^{\sigma(a,b)} g(b) \mid b \in B\} \\ f^{\downarrow}(b) &= \inf\{R(a,b) \nwarrow_{\sigma(a,b)} f(a) \mid a \in A\} \end{aligned}$$

for all $g \in L_2^B$, $f \in L_1^A$. These two operators form a Galois connection [13]. The notion of concept can be defined as usual: A multi-adjoint concept is a pair $\langle g, f \rangle$ satisfying that $g \in L_2^B$, $f \in L_1^A$ and $g^{\uparrow} = f$, $f^{\downarrow} = g$, with $(^{\uparrow}, ^{\downarrow})$ being the Galois connection defined above. The fuzzy subsets g and f are usually known as the extension and intension of a concept, respectively.

Definition 5. *Given a multi-adjoint frame* $(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$ *and a multi-adjoint context* (A, B, R, σ)*, the multi-adjoint concept lattice associated with them is the set*

$$\mathscr{M} = \{\langle g, f \rangle \mid g \in L_2^B, f \in L_1^A \text{ and } g^{\uparrow} = f, f^{\downarrow} = g\}$$

in which the ordering is defined by $\langle g_1, f_1 \rangle \preceq \langle g_2, f_2 \rangle$ *if and only if* $g_1 \preceq_2 g_2$ *(or equivalently,* $f_2 \preceq_1 f_1$*).*

In the following, we will present a characterization of the join-irreducible elements of a multi-adjoint concept lattice. A similar result with respect to the meet-irreducible elements was introduced in [5]. First and foremost, we need to consider a multi-adjoint concept lattice associated with a multi-adjoint frame and a multi-adjoint context, where L_1, L_2, P, A and B are finite, and the following specific family of fuzzy subsets of L_2^B.

Definition 6. *For each $b \in B$, the fuzzy subsets of attributes $\phi_{b,y} \in L_2^B$ defined, for all $y \in L_2$, as*

$$\phi_{b,y}(b') = \begin{cases} y & \text{if } b' = b \\ 0 & \text{if } b' \neq b \end{cases}$$

will be called fuzzy-objects. *The set of all fuzzy-objects will be denoted as $\Psi = \{\phi_{b,y} \mid b \in B, y \in L_2\}$.*

The next theorem establishes that the join-irreducible elements of a concept lattice are uniquely generated by fuzzy-objects. In addition, they cannot be expressed as supremum of elements that are less than them. A dual result related to the meet-irreducible elements of a concept lattice was presented in [5].

Theorem 1 ([5])**.** *The set of join-irreducible elements of $\mathscr{M}$, $J_F(\mathscr{M})$, is formed by the pairs $\langle \phi_{b,y}^{\uparrow\downarrow}, \phi_{b,y}^{\uparrow} \rangle$ in $\mathscr{M}$, with $b \in B$ and $y \in L_2$, such that*

$$\phi_{b,y}^{\uparrow} \neq \bigwedge \{\phi_{b_i,y_i}^{\uparrow} \mid \phi_{b_i,y_i} \in \Psi, \phi_{b,y}^{\uparrow} \prec_1 \phi_{b_i,y_i}^{\uparrow}\}$$

and $\phi_{b,y}^{\uparrow} \neq f_{\top_1}$, where Ψ denotes the set of all fuzzy-objects, $\top_1$ is the maximum element in L_1 and $f_{\top_1} : A \to L_1$ is the fuzzy subset defined as $f_{\top_1}(a) = \top_1$, for all $a \in A$.

In this paper, we are also interested in considering a multi-adjoint concept lattice framework enriched with hedges. Given two arbitrary hedges $*_A$ and $*_B$, if we define the concept-forming operators ${}^{\uparrow *_B} : L_2^B \to L_1^A$ and ${}^{\downarrow *_A} : L_1^A \to L_2^B$ as:

$$g^{\uparrow *_B}(a) = \inf\{R(a,b) \swarrow^{\sigma(a,b)} g(b)^{*_B} \mid b \in B\}$$
$$f^{\downarrow *_A}(b) = \inf\{R(a,b) \nwarrow_{\sigma(a,b)} f(a)^{*_A} \mid a \in A\}$$

for all $g \in L_2^B$, $f \in L_1^A$, these operators do not form a Galois connection in general [1]. Taking into account the definition of the identity hedge, in order to simplify the notation, we will write $({}^{\uparrow *}, {}^{\downarrow})$ instead of $({}^{\uparrow *_B}, {}^{\downarrow *_A})$.

The context associated with the previous concept-forming operators will be denoted as (A, B^*, R, σ) and the concepts will be defined as the pairs $\langle g, f \rangle$ satisfying that $g^{\uparrow *} = f$, $f^{\downarrow} = g$, for all $g \in L_2^B$, $f \in L_1^A$.

Definition 7. *The multi-adjoint concept lattice with hedges associated with a multi-adjoint frame $(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$ and a context (A, B^*, R, σ) is the set*

$$\mathscr{M}_* = \{\langle g, f \rangle \mid g \in L_2^B, f \in L_1^A \text{ and } g^{\uparrow *} = f, f^{\downarrow} = g\}$$

in which the ordering is defined by $\langle g_1, f_1 \rangle \preceq \langle g_2, f_2 \rangle$ if and only if $g_1 \preceq_2 g_2$ (or equivalently, $f_2 \preceq_1 f_1$).

 ISBN: 978-84-608-6082-2

4 Validity of fuzzy attribute implications

As we mentioned at the beginning of the paper, our goal focuses on the survey of attribute implications making use of the powerful advantages provided by the multi-adjoint environment. Namely, our contribution will allow us to complement the study on attribute implications in multi-adjoint concept lattices presented in [12].

From now on, the multi-adjoint frame $(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$ and the contexts (A, B, R, σ) and (A, B^*, R, σ) will be fixed. The corresponding multi-adjoint lattices will be denoted as: $(\mathscr{M}, \preceq)$ and $(\mathscr{M}_*, \preceq)$, respectively.

Now, we will present a generalization of the classical notion of fuzzy implication considering adjoint triples.

Definition 8. *Given an adjoint triple $(\&, \swarrow, \nwarrow)$ with respect to L_1, L_2, P and two fuzzy subsets of attributes $f_1, f_2 \in L_1^A$,* the degree in which f_1 is included in f_2 *is defined as*

$$S(f_1, f_2) = \inf\{f_2(a) \nwarrow f_1(a) \mid a \in A\}$$

A truth-stressing hedge is considered in the definition of the degree in which a fuzzy implication is valid.

Definition 9. *Given an adjoint triple $(\&, \swarrow, \nwarrow)$ with respect to L_1, L_2, P and three fuzzy subsets of attributes $f_1, f_2, f_3 \in L_1^A$, the degree in which a fuzzy implication $f_2 \Leftarrow f_1$ is valid in f_3 is*

$$\|f_2 \Leftarrow f_1\|_{f_3} = S(f_2, f_3) \nwarrow S(f_1, f_3)^*$$

where $*$ is a truth-stressing hedge.

This notion is extrapolated to a set of fuzzy subsets of attributes $\mathscr{F} \subset L_1^A$, defining the degree in which the fuzzy implication $f_2 \Leftarrow f_1$ is valid in $\mathscr{F}$ as

$$\|f_2 \Leftarrow f_1\|_{\mathscr{F}} = \bigwedge_{f \in \mathscr{F}} S(f_2, f) \nwarrow S(f_1, f)^*$$

The following definition introduces the notion of validity of a fuzzy implication on a multi-adjoint concept lattice.

Definition 10. *Let $(\&, \swarrow, \nwarrow)$ be an adjoint triple with respect to L_1, L_2, P and $f_1, f_2 \in L_1^A$ two fuzzy subsets of attributes,* the degree in which the fuzzy implication $f_2 \Leftarrow f_1$ is valid in the multi-adjoint concept lattice $\mathscr{M}_*$ *is the truth degree:*

$$\|f_2 \Leftarrow f_1\|_{\mathscr{M}_*} = \|f_2 \Leftarrow f_1\|_{Int(\mathscr{M}_*)}$$

where $Int(\mathscr{M}_*)$ is the set of intensions of the concepts in $(\mathscr{M}_*, \preceq)$.

The following result reveals that the degree in which a fuzzy implication is valid in a multi-adjoint concept lattice is obtained computing only the degree of validity on the set of join-irreducible elements of the lattice [5]. As a consequence, we do not need to calculate the degree of validity on the whole set of intensions of the concepts of the lattice, which imply an important reduction in the number of calculations.

Theorem 2. *Let $f_1, f_2 \in L_1^A$ be fuzzy subsets of attributes. Then,*

$$\|f_2 \Leftarrow f_1\|_{Int(\mathscr{M}_*)} = \|f_2 \Leftarrow f_1\|_{Int(J_F(\mathscr{M}_*))}$$

where $Int(J_F(\mathscr{M}_*))$ denotes the set of intensions of the join-irreducible elements of the lattice associated with (A, B^*, R, σ).

Finally, in order to clarify Theorem 2, we will include an illustrative example that computes the validity of a particular fuzzy attribute implication and shows that the number of operations is noticeably reduced.

Example 1. *It will be considered the framework $(L, \preceq, \&_P)$, where $L = [0,1]_{10}$ is the regular partitions of $[0,1]$ in 10 pieces and $\&_P$ is the discretization of the product conjunctor, see [4] for more details. The fixed context is (A, B^*, R, σ), with $A = \{a_1, a_2, a_3, a_4, a_5\}$, $B = \{b_1, b_2, b_3\}$, $R\colon A \times B \to L$ given by the Table 1, the hedge $*$ is the identity and σ is constantly $\&_P$.*

The concept lattice related to this framework and this context is composed by 33 concepts whose intensions are:

$Int(C_0) = \{1.0/a_1, 1.0/a_2, 1.0/a_3, 1.0/a_4, 1.0/a_5\}$
$Int(C_1) = \{1.0/a_1, 1.0/a_3, 1.0/a_4, 1.0/a_5\}$
$Int(C_2) = \{1.0/a_2\}$
$Int(C_3) = \{0.8/a_1, 0.8/a_3, 0.8/a_4, 1.0/a_5\}$
$Int(C_4) = \{0.5/b_1, 0.5/b_2\}, \{1.0/a_1, 1.0/a_3, 1.0/a_5\}$
$Int(C_5) = \{0.7/a_1, 0.7/a_3, 0.7/a_4, 1.0/a_5\}$
$Int(C_6) = \{0.8/a_1, 0.8/a_3, 1.0/a_5\}$
$Int(C_7) = \{0.6/a_1, 0.6/a_3, 0.6/a_4, 1.0/a_5\}$
$Int(C_8) = \{0.7/a_1, 0.7/a_3, 1.0/a_5\}$
$Int(C_9) = \{0.5/a_1, 0.5/a_3, 0.5/a_4, 1.0/a_5\}$
$Int(C_{10}) = \{0.6/a_1, 0.6/a_3, 1.0/a_5\}$
$Int(C_{11}) = \{0.5/a_1, 0.5/a_3, 1.0/a_5\}$
$Int(C_{12}) = \{0.5/a_1, 0.5/a_3, 0.8/a_5\}$
$Int(C_{13}) = \{0.5/a_1, 0.5/a_3, 0.7/a_5\}$
$Int(C_{14}) = \{0.5/a_1, 0.5/a_3, 0.6/a_5\}$
$Int(C_{15}) = \{0.5/a_1, 0.5/a_3, 0.5/a_5\}$
$Int(C_{16}) = \{\}$
$Int(C_{17}) = \{0.6/a_1, 0.6/a_3, 0.8/a_5\}$
$Int(C_{18}) = \{0.6/a_1, 0.6/a_3, 0.7/a_5\}$
$Int(C_{19}) = \{0.6/a_1, 0.6/a_3, 0.6/a_5\}$
$Int(C_{20}) = \{0.6/a_1, 0.5/a_3, 0.5/a_5\}$
$Int(C_{21}) = \{0.7/a_1, 0.7/a_3, 0.8/a_5\}$
$Int(C_{22}) = \{0.7/a_1, 0.7/a_3, 0.7/a_5\}$
$Int(C_{23}) = \{0.7/a_1, 0.6/a_3, 0.6/a_5\}$
$Int(C_{24}) = \{0.7/a_1, 0.5/a_3, 0.5/a_5\}$
$Int(C_{25}) = \{0.8/a_1, 0.8/a_3, 0.8/a_5\}$
$Int(C_{26}) = \{0.8/a_1, 0.7/a_3, 0.7/a_5\}$
$Int(C_{27}) = \{0.8/a_1, 0.6/a_3, 0.6/a_5\}$
$Int(C_{28}) = \{0.8/a_1, 0.5/a_3, 0.5/a_5\}$
$Int(C_{29}) = \{1.0/a_1, 0.8/a_3, 0.8/a_5\}$
$Int(C_{30}) = \{1.0/a_1, 0.7/a_3, 0.7/a_5\}$
$Int(C_{31}) = \{1.0/a_1, 0.6/a_3, 0.6/a_5\}$
$Int(C_{32}) = \{1.0/a_1, 0.5/a_3, 0.5/a_5\}$

R	b_1	b_2	b_3
a_1	0.5	1	0
a_2	0	0	1
a_3	0.5	0.5	0
a_4	0.5	0	0
a_5	1	0.5	0

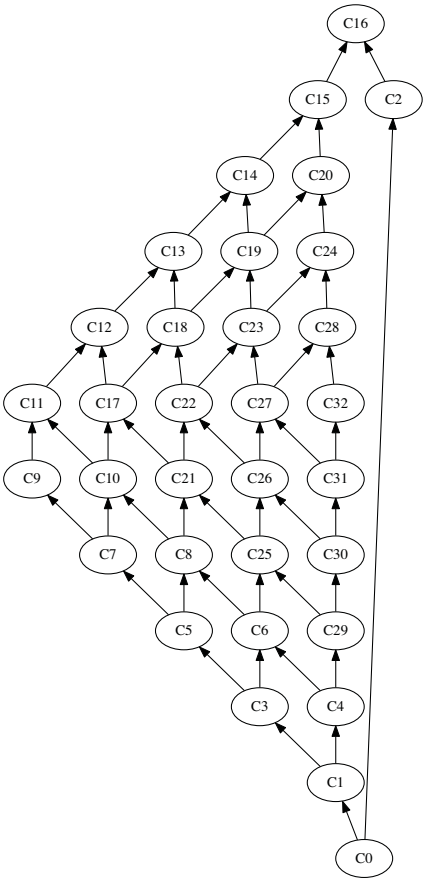


Figure 1: Relation R and Hasse diagram of Example 1.

From Theorem 1, we obtain that the set of join-irreducible elements associated with the concept lattice is:

$$J_F(\mathscr{M}_*) = \{C_1, C_2, C_3, C_4, C_5, C_7, C_9, C_{29}, C_{30}, C_{31}, C_{32}\}$$

Now, we will consider the fuzzy subsets of attributes $f_1 = \{0.5/a_1, 0.5/a_3, 0.5/a_5\}$, $f_2 = \{0.1/a_5\}$ *and we will compute the degree in which the fuzzy implication* $f_2 \Leftarrow f_1$ *is valid in* $Int(\mathscr{M}_*)$. *Instead of using all intensions of the concept lattice related to the context* (A, B^*, R, σ), *by Theorem 2, we can reduce the number of computations. We will only consider the intensions of the join-irreducible elements of the concept lattice associated with* (A, B^*, R, σ), *which are* $|J_F(\mathscr{M}_*)| = 11 < 33$. *Applying Definition 9, we have that:*

$$\begin{aligned} \|f_2 \Leftarrow f_1\|_{Int(J_F(\mathscr{M}_*))} &= \bigwedge\{\|f_2 \Leftarrow f_1\|_{Int(C)} \mid C \in J_F(\mathscr{M}_*)\} \\ &= \bigwedge\{(S(f_2, Int(C)) \nwarrow S(f_1, Int(C)^*) \mid C \in J_F(\mathscr{M}_*)\} = 1 \end{aligned}$$

where Int(C) denotes the intension of a concept.

5 Conclusions and future work

We have proven that the computation of the validity of attribute implications can be carried out by means of join-irreducible elements of a multi-adjoint concept lattice. This procedure can considerably reduce the number of operations if we compare it with the computations from the whole set of intensions of the concept lattice. Moreover, we have obtained an interesting property which relates the boundary conditions of an adjoint triple with the monotonicity of a truth-stressing hedge. This property has played an important role in the proof of the main result presented in this paper.

As a future work, we will study more properties of the attribute implications in multi-adjoint framework. In addition, we are interested in performing experimental surveys about the generation of attribute implications.

References

[1] R. Bělohlávek. Sup-t-norm and inf-residuum are one type of relational product: Unifying framework and consequences. *Fuzzy Sets and Systems*, 197:45–58, 2012.

[2] R. Bělohlávek, P. Cordero, M. Enciso, A. Mora, and V. Vychodil. Automated prover for attribute dependencies in data with grades. *International Journal of Approximate Reasoning*, 70:51 – 67, 2016.

[3] P. Cordero, M. Enciso, A. Mora, and M. Ojeda-Aciego. Computing left-minimal direct basis of implications. In M. Ojeda-Aciego and J. Outrata, editors, *CLA*, volume 1062 of *CEUR Workshop Proceedings*, pages 293–298. CEUR-WS.org, 2013.

[4] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. A comparative study of adjoint triples. *Fuzzy Sets and Systems*, 211:1–14, 2013.

[5] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. Attribute reduction in multi-adjoint concept lattices. *Information Sciences*, 294:41 – 56, 2015.

[6] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. Multi-adjoint algebras versus extended-order algebras. *Applied Mathematics & Information Sciences*, 9(2L):365–372, 2015.

[7] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. Multi-adjoint algebras versus non-commutative residuated structures. *International Journal of Approximate Reasoning*, 66:119–138, 2015.

[8] C. V. Glodeanu. Knowledge discovery in data sets with graded attributes. *International Journal of General Systems*, 45(2):232–249, 2016.

[9] P. Hájek. On very true. *Fuzzy Sets and Systems*, 124(3):329 – 333, 2001.

[10] J. Hill, H. Walkington, and D. France. Graduate attributes: implications for higher education practice and policy. *Journal of Geography in Higher Education*, 40(2):155–163, 2016.

[11] T. Kuhr and V. Vychodil. Fuzzy logic programming reduced to reasoning with attribute implications. *Fuzzy Sets and Systems*, 262:1 – 20, 2015.

[12] V. Liñeiro-Barea, J. Medina, and I. Medina-Bulo. Towards generating fuzzy rules via fuzzy formal concept analysis. In J. Kacprzyk, L. Koczy, and J. Medina, editors, *7th European Symposium on Computational Intelligence and Mathematices (ESCIM 2015)*, pages 60–65, 2015.

[13] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. Formal concept analysis via multi-adjoint concept lattices. *Fuzzy Sets and Systems*, 160(2):130–144, 2009.

[14] J. Medina, M. Ojeda-Aciego, and P. Vojtáš. Multi-adjoint logic programming with continuous semantics. In *Logic Programming and Non-Monotonic Reasoning, LPNMR'01*, pages 351–364. Lecture Notes in Artificial Intelligence 2173, 2001.

[15] J. Medina, M. Ojeda-Aciego, and P. Vojtáš. Similarity-based unification: a multi-adjoint approach. *Fuzzy Sets and Systems*, 146:43–62, 2004.

[16] J. Rodríguez-Jiménez, P. Cordero, M. Enciso, and A. Mora. Negative attributes and implications in formal concept analysis. *Procedia Computer Science*, 31:758 – 765, 2014.

[17] V. Vychodil. Computing sets of graded attribute implications with witnessed non-redundancy. *Information Sciences*, 351:90 – 100, 2016.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On applying a parallel Teaching-Learning-Based optimization procedure for automatic heliostat aiming

N.C. Cruz[1], J.L. Redondo[1], J.D. Álvarez[1], M. Berenguel[1] and P.M. Ortigosa[1]

[1] *Dpt. of Informatics, University of Almería, CeiA3, Spain*

emails: ncalvocruz@ual.es, jlredondo@ual.es, jhervas@ual.es, beren@ual.es, ortigosa@ual.es

Abstract

The operative configuration of the heliostat field of solar central receiver plants is a vital part of their controlling tasks. The subset of active heliostats must be carefully configured to set the operational state as desired while also avoiding dangerous flux distributions and radiation peaks over the receiver surface. In this context, a general and automatic aiming methodology is being developed by the authors of this work. However, the mathematical formulation of this problem leads to a complex large-scale optimization problem in which every active heliostat requires a certain two-dimensional aiming point over the receiver. In this work, the possibility of applying TLBO, a population based large-scale optimizer, is studied. Considering the potential computational costs of this task, a preliminary parallel version of TLBO has been developed. The application of this method to perform a large exploration of the search-space, in a high-performance computing environment, is described. The parallelization of the algorithm turns out to be quite useful to accelerate the procedure for the problem at hand. Therefore, the possibility of including additional steps to the method remains feasible.

Key words: Parallel computing, large-scale optimization, TLBO, Heliostat aiming

1 Introduction

Solar Central Receiver Systems, SCRS in what follows, are power generation facilities based on the exploitation of solar energy by concentrating the incident radiation. In general terms, considering the scope of this work, they are formed by a large group of high-reflectance orientable mirrors and a radiation receiver on the top of a tower. The mirrors, which are

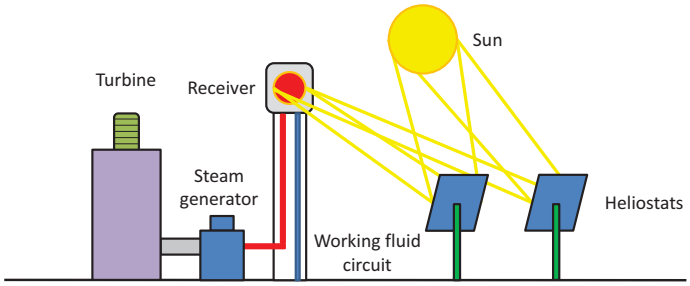


Figure 1: Scheme of a solar tower power plant.

called 'heliostats', track the apparent movement of the Sun through the day to concentrate the incident solar radiation over the receiver. Consequently, there is a very high radiation density over its surface. This power is then transferred to a working fluid which is in circulation inside the receiver. After increasing its temperature, this fluid can be finally used in a classic thermodynamic cycle for electric power generation. In Fig. 1, an illustrative schema of this kind of facilities is shown. Stability of production and operative efficiency due to the maturity of the underlying technologies are key aspects of this kind of power facilities. The interested reader is referred to [2, 9] for further information about them.

Controlling the flux distribution formed by the heliostat field over the receiver surface is of major importance to avoid dangerous temperature gradients, thermal stress and premature aging of its components [1, 3, 7, 11]. This is a key factor for increasing the operative life of the receiver, what has a direct influence on the production costs of STTP as highlighted in [7]. Considering that the heliostat field is usually formed by hundreds (occasionally thousands) of heliostats, the definition of the active subset of them, as well as their specific aiming point over the receiver according to the desired flux distribution, leads to face a very complex multi-staggered problem. In [3, 7], this problem is addressed by a pre-defined set of heliostats and possible aiming points, looking for an homogeneous flux distribution, with good results. However, in the context of this work, a generalization of their approach is being developed by trying to automatically configure the whole field for a given instant of time (i.e., solar position) and a desired flux distribution to achieve. This methodology would even include the possibility of disabling unnecessary heliostats to replicate the given reference (as heliostat fields are commonly oversized to face unfavorable operating conditions such as cloudy days). In any case, that step is out of the scope of the present paper and it will be assumed that the selected heliostats are already known. At this point, good overall results are obtained by applying gradient-based local optimizers to define their corresponding aiming point. Unfortunately, these approaches have a local scope and the objective function is known to have multiple local optima. It is intended to add

a computationally efficient global optimizer to the process to get a wide perspective of the problem. An existing large-scale-oriented population-based global optimizer, the Teaching-Learning-Based algorithm (TLBO) [5], will be considered. Consequently, the deployment of this method in a high-performance computing environment is the main aim of this work.

In Section 2, the problem at hand is formally described. Then, in Section 3, the TLBO algorithm is exposed. In Section 4, the selected parallelization strategy for the algorithm is commented. Finally, experimentation and results are shown and conclusions are drawn in Sections 5 and 6 respectively.

2 Problem definition

In order to model the present problem, it is necessary to define the target flux distribution to achieve, F. It is a matrix of size $Y \times X$, where X and Y are referred to the number of rows and columns of the matrix respectively. All the elements of F are known in both position and magnitude (flux density) as it is part of the information of the problem. This matrix can be seen as a 'picture' of the flux distribution to replicate over the receiver, which is considered as a flat rectangle, with the active heliostats. Its dimension Y is linked to the vertical of the receiver plane, the direction from the plane towards the zenith in a three-dimensional Cartesian system. Similarly, its dimension X is linked to the horizontal direction of the receiver plane along the West-East direction. In this context, the discretization axes are also known as vectors $Y' = y_0, \cdots, y_Y$ and $X' = x_0, \cdots, x_X$ whose length is Y and X respectively. Consequently, every element of the matrix is referred to a particular zone of the receiver, whose surface is intrinsically discretized by the step used when defining F.

In relation to the active heliostats, it is an ordered set $H = \{h_1, h_2, \cdots, h_T\}$ with cardinality T. Every heliostat h_i projects a certain flux distribution f_{h_i} over the receiver when it is operative, which is a known bi-dimensional continuous function of the radiation density. It is defined as a bi-dimensional Gaussian density function as shown in Eq. 1

$$f_{h_i}(x,y) = \frac{P}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right)} \tag{1}$$

where x and y are the coordinates on the plane defined by the receiver rectangular aperture in its X and Y dimensions respectively, P is the power contribution of the heliostat h_i over the receiver, ρ is the correlation between x and y, σ_x and σ_y are the standard deviation along x and y respectively and μ_x and μ_y, which are the mean in the Gaussian probability function, define the central point of the flux distribution, i.e. the aiming point of the heliostat h_i. This approach is similar to the one selected by [3, 7], where a specific circular Gaussian density function is applied according to the HFLCAL model [8]. All the parameters that define the shape of the flux distribution of every heliostat, P, ρ, σ_x and σ_y, are known (from detailed simulation and curve-fitting procedures) while its central point (μ_x, μ_y) should be

 ISBN: 978-84-608-6082-2

determined to replicate the reference. Once every heliostat has a certain aiming point, the configuration vector of the field c can then be defined by concatenating the pair of aiming coordinates of every heliostat. Finally, the obtained flux distribution, F_c^*, is formed then by the convolution of the particular flux distribution of each one, over the receiver plane, along the discretization axises.

Taking into consideration the previous definitions, a $2T$-dimensional minimization problem can be formulated as the accumulation of the square difference, at every discretization point, between the matrices F and F_c^* as shown in Eq. 2.

$$min\, O = min \sum_{x=x_0}^{x_X} \sum_{y=y_0}^{y_Y} \left(F(x,y) - F_c^*(x,y)\right)^2 \tag{2}$$

3 Teaching-Learning-Based Optimization

Teaching-Learning-Based Optimization, TLBO, is an stochastic population-based global optimizer presented in [5]. The algorithm models the behavior of a class of students, who form the population of candidate solutions. They are progressively improved by simulating both the teaching process of the teacher of the class and the interaction between the students. This algorithm is mainly characterized by its performance and its virtual lack of specific parameters, as only the population size and the number of cycles need to be specified by the user. Although the original form of the algorithm for continuous non-constrained problems has been selected for this work, the interested reader is referred to [6] for further information about it, its versions and applications. The linked works of [4] and [10] should be also examined by the reader interested in TLBO.

Students, i.e. candidate solutions, are defined as N-dimensional vectors, where N is the number of dimensions of the studied problem. Every dimension is considered as a 'subject' in the context of TLBO, and the natural representation of the class is a $P \times N$ matrix in which P is the population size. The quality of the class is assumed to follow a normal distribution whose average value must be increased by academic interaction. Consequently, the value of students at every subject is altered along the cycles in order to improve the overall average as desired. To achieve it, the plain TLBO algorithm relies on two stages per cycle, the Teacher and the Learners phases [5], which are summarized next for a certain cycle k from a minimization perspective.

3.1 Teacher Phase (TS)

At this step, the average value per subject, i.e. per column, is calculated from the current population. An N-dimensional vector, M, is obtained. Then, the best student, whose value according to the objective function should be the minimum, is promoted to become the teacher of the cycle. A random integer in $[1, 2]$ called 'Teaching Factor', T_F, is generated

as an overall weighting factor of its teaching capabilities. After that, an N-dimensional random vector r of real values in $[0, 1]$ is generated to model the skills of the teacher at teaching every subject and the aptitudes of the students while learning them. With this information, an N-dimensional global shifting vector DM is computed according to Eq. 3, where T is the content of the teacher as a candidate solution. It must be noted that Eq. 3 is referred to dimensional element-by-element operations. Finally, DM is added as a vector to every existing individual. However, only improved students, after the evaluation of the objective function, are kept, while the change is discarded otherwise.

$$DM = r(T - T_F M) \tag{3}$$

3.2 Learner Phase (LS)

At this step, an N-dimensional random vector r of real values in $[0, 1]$ is generated to model the advancing possibilities at every subject in a similar way as done for the TS. Then, every final individual i from the previous step is randomly paired with another one j different from itself. After that, the individual i is shifted from its current position depending on whether i has a better value of the objective function than j or not as noted in Eq. 4, where the dimensional element-by-element operation scheme is maintained. Finally, the individual i is only updated when its value has been improved after the change.

$$Student_i = \begin{cases} Student_i + r(Student_j - Student_i), & \text{if } j \text{ better than } i \\ Student_i + r(Student_i - Student_j), & \text{if } i \text{ better than } j \end{cases} \tag{4}$$

At this point, the population of the next cycle would have been defined. However, an additional step not commented in [5] may should have been included to remove duplicate solutions, what is highlighted in [4] and addressed in [10]. This procedure is expected to look for equal solutions and to randomly re-initialize a dimension of one of them, what requires its re-evaluation as candidate solution.

4 Parallelization strategy

Considering the described structure of TLBO, and assuming a computationally demanding objective function, the underlying iterative structures linked to both TS and LS (updating and studying every student) could be assigned to different execution units. By proceeding this way, the whole procedure would be the same but, at every stage, the management of the available individuals would be distributed between execution units. Consequently, the required evaluations of the objective function would be directly shared between the available execution units. This is the selected approach for the present work in a thread-based environment. Its efficiency is linked to the number of individuals and the intrinsic

cost of the evaluation of the objective function. However, it does not depend on the number of cycles.

Finally, it is important to mention that for extremely large populations and/or hard objective functions, the previous strategy could be easily generalized. Particularly, the population could be divided in different subsets that would be internally altered and evaluated also in parallel. It would be suitable for a hybrid process-thread-based environment when being able to amortize the communication costs.

5 Experimentation and results

A sample instance of the defined problem aims to form an homogeneous flat form over the receiver. Consequently, matrix F is formed by a single and replicated value: 80 kW/m^2 over a 6×6 meters receiver. The subset of heliostats to activate is adequately selected by our current procedure, which selects 110 heliostats to deploy. In this context, the developed parallel TLBO implementation will be launched with different configurations to check its computational performance. It has been implemented in C with OpenMP directives for threading purposes.

The execution platform is a cluster node featuring an Intel Xeon E5 2650v2 with 16 cores and 128 GB RAM. The number of cycles of TLBO has been fixed to 150 after preliminary adjustment. In relation to the number of individuals, what sets the direct computational load of every cycle, it has been configured to be 50, 100, 200 and 400. Considering the hardware, the number of threads has been fixed to 2, 4, 8 and 16 for all cases.

In Fig. 2, the speedup achieved with the parallel version of TLBO is shown for the different population sizes. Additionally, a black dotted line represents the theoretical linear speedup. These results have been averaged after five executions. As can be seen, the speedup is almost linear for all the instances. In fact, with 2 and 4 threads, it could be considered linear for all the population sizes. The peak of performance is achieved with the largest population and 16 active threads, where the speedup is 14.10. However, as expected, it is slightly worse when the population size is reduced and the number of active threads is too high. In other words, the speedup is progressively separated from linearity when the ratio between active threads and individuals is reduced. Additionally, considering the monotonic ascending tendency in all cases, the scalability of the process can be also highlighted.

Finally, in Fig. 3, the corresponding flux distribution of the best solution found by TLBO is shown. It has been obtained with a population of 400 individuals along 150 cycles of search. Its top is relatively flat and near 80 kW/m^2 as intended. Consequently, the algorithm is not simply compatible with an efficient parallel execution, but its results seem to be promising to be used as the initial point of further local gradient-based optimizers.

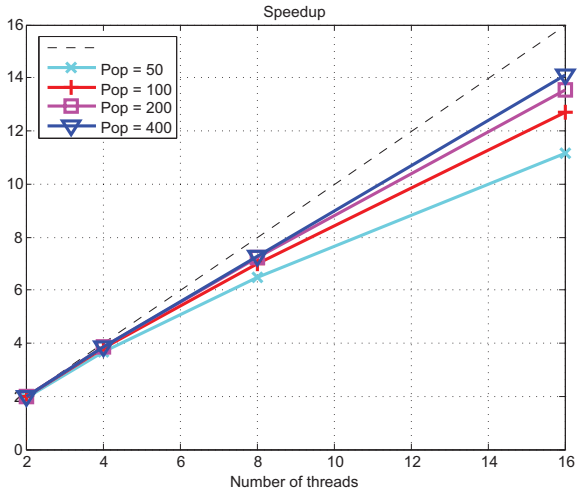


Figure 2: Achieved speedup with the thread-based parallel version of TLBO.

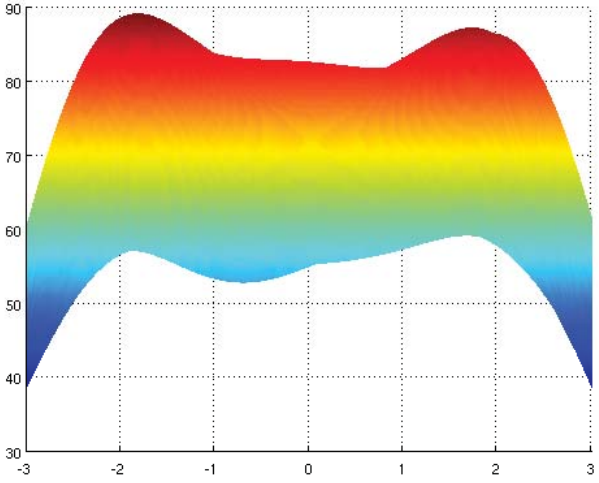


Figure 3: X-Z plane of the obtained solution by TLBO when reproducing a flat distribution.

 ISBN: 978-84-608-6082-2

6 Conclusions and future work

In this work, the problem of defining the aiming points of a set of heliostats has been presented. Then, it has been formalized from a mathematical perspective as an unconstrained large-scale minimization problem. It is known to have numerous local optima and a complexity which is increased with the number of active heliostats. Therefore, an existing population-based global optimizer, TLBO, has been selected to be studied for efficient vast explorations of the search-space. Considering the computational cost of the objective function and the necessity of handling large populations, an OpenMP-based TLBO implementation has been developed. It shows an almost-linear speedup and an scalable behavior with acceptable results. Consequently, it seems to be adequate to be used as the global guide of further local optimization stages.

As future work, taking into account the positive results, the inclusion of the parallel TLBO optimizer in our automatic heliostat aiming procedure will be considered in depth. Additionally, the possibility of deploying a hybrid parallel version of TLBO that can be executed in different nodes of a cluster will be addressed.

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Economy and Competitiveness (TIN2015-66680-C2-1-R and ENERPRO DPI 2014-56364-C2-1-R), Junta de Andalucía (P11-TIC7176 and P12-TIC301). Nicolás Calvo Cruz is supported by a FPU Fellowship from the Spanish Ministry of Education. Juana López Redondo and José Domingo Álvarez Hervás are fellows of the Spanish 'Ramón y Cajal' contract program, co-financed by the European Social Fund.

References

[1] M. Carasso and M. Becker. Solar Thermal Central Receiver Systems: Performance evaluation standards for solar central receivers (Vol. 3). *Springer Verlag*, 1991.

[2] O. Behar, A. Khellaf and K. Mohammedi. A review of studies on central receiver solar thermal power plants. *Renewable and Sustainable Energy Reviews*, 23:12–39, 2013.

[3] S.M. Besarati, D.Y. Goswami and E.K. Stefanakos. Optimal heliostat aiming strategy for uniform distribution of heat flux on the receiver of a solar power tower plant. *Energy Conversion and Management*, 84:234–243, 2014.

[4] M. Crepinsek, S.H. Liu and L. Mernik. A note on teachinglearning-based optimization algorithm. *Information Sciences*, 212:79-93, 2012.

[5] R.V. Rao, V.J. Savsani and D.P. Vakharia. Teaching-learning-based optimization: an optimization method for continuous non-linear large scale problems. *Information Sciences*, 183(1):1-15, 2012.

[6] R.V. Rao. *Teaching Learning Based Optimization Algorithm: And Its Engineering Applications. Springer*. 2015.

[7] A. Salomé, F. Chhel, G. Flamant, A. Ferrière and F. Thiery. Control of the flux distribution on a solar tower receiver using an optimized aiming point strategy: Application to THEMIS. *Solar Energy*, 94:352–366, 2013.

[8] P. Schwarzbözl, M. Schmitz and R. Pitz-Paal. Visual HFLCAL - A Software Tool for Layout and Optimisation of Heliostat Fields. In SolarPACES'09, 2009.

[9] W. Stine and M. Geyer. Power from the Sun, 2001. Public website, available from http://powerfromthesun.net/book.html (Last access: May 10, 2016)

[10] G. Waghmare. Comments on 'A note on teaching-learning-based optimization algorithm'. *Information Sciences*, 229:159-169, 2013.

[11] C.J. Winter, R.L. Sizmann and L.L. Vant-Hull. Solar power plants: fundamentals, technology, systems, economics. *Springer-Verlag New York*, 1991.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Empirical Modeling: an Auto-tuning Method for Linear Algebra Routines on CPU+multiGPUs Platforms

Javier Cuenca[1], Luis P. García[2], Domingo Giménez[3] and Francisco J. Herrera[3]

[1] *Department of Engineering and Technology of Computers, University of Murcia, Spain*

[2] *Service of Support to Technological Research, Technical University of Cartagena, Spain*

[3] *Department of Computing and System, University of Murcia, Spain*

emails: jcuenca@um.es, Luis.Garcia@sait.upct.es, domingo@um.es, franciscojose.herrera@um.es

Abstract

Scientific software must be optimized, by hand or automatically, for current hybrid computational systems, which are composed normally of a multicore CPU and several manycore coprocessors of different types. This work analyses how to apply an auto-tuning technique, based on execution time modeling, to improve routine performance on hybrid platforms composed of a CPU plus several GPUs. The final goal is to obtain a balanced assignation of the work to the computing components in the system and to decide the best number of CPU threads. Experimental results are satisfactory with different numbers and kinds of GPUs.

Key words: auto-tuning; parallel linear algebra; manycore; hybrid programming; heterogeneous computing

1 Introduction

Among the most important bases that support the scientific and engineering software are basic BLAS-type matrix routines. Therefore, the improvement in the performance of scientific codes is achieved in many cases by the efficient use of these routines.

There are important research projects whose main goal is the optimization of linear algebra routines in computational systems of different characteristics [1, 2, 3, 4, 5]. Traditional approaches for homogeneous platforms are adapted to heterogeneous or dynamic

 ISBN: 978-84-608-6082-2

systems [6, 7, 8, 9]. Nowadays CPU+coprocessor systems occupy the most important position among high performance computing systems, so it is necessary to adapt the existing auto-tuning techniques for these platforms.

Several BLAS implementations exist for multicore CPUs (vendors implementations: Intel MKL [10], IBM ESSL [11], etc., or free implementations: ATLAS [5], Goto BLAS [12], etc.) and for manycore GPUs (CULA Tools [13], CUBLAS [14] and MAGMA [15]). The CPU and GPU implementations used in this work are MKL and CUBLAS, but the same methodology described here can be applied with other libraries.

This paper describes an auto-tuning technique whose objective is to reduce the execution time of linear algebra routines that are executed on hybrid systems composed of a multicore CPU and several manycore GPUs. This optimization engine has to take two important decisions: the load balance between the CPU and the GPUs, and how many CPU threads to generate. In previous works we considered two different auto-tuning methods to obtain a balanced distribution of the tasks between the computing resources according to the machine characteristics: an experimental method (guided search) and a mixed theoretical-experimental method (empirical modeling). The guided search method was analyzed in [16] and the empirical modeling in [17], in both cases for NUMA platforms. Their extension to CPU+GPU platforms is described in [18] (guided search) and [19, 20] (empirical modeling). After that, the guided search was also adapted to CPU+multiGPUs in [21]. Here, the empirical modeling is adapted to hybrid platforms composed of a multicore CPU plus multiple GPUs of different characteristics.

The rest of the paper is organized as follows. Section 2 describes the hybrid matrix multiplication kernel for multicore+multiGPU platforms. Section 3 introduces the auto-tuning methodology and explains how to use it for this basic kernel in this type of hybrid systems. The experimental results are shown in Section 4. Section 5 concludes and outlines possible research directions.

2 Hybrid matrix multiplication kernel

Simultaneous work of the CPU cores and the coprocessors is considered in order to optimize the matrix multiplication. If only one coprocessor is available, the multiplication $C = \alpha AB + \beta C$ can be expressed as $C = \alpha(AB_1 + AB_2) + \beta(C_1 + C_2)$, with the multiplication $\alpha AB_1 + \beta C_1$ assigned to the coprocessor and $\alpha AB_2 + \beta C_2$ to the CPU. An optimum split of the matrix would balance the work assigned to the CPU and to the coprocessor [22, 23]. The computations carried out in the two computational systems overlap. Computation and communication should overlap for maximum performance, and so the data transfers between the multicore and the coprocessor should be performed asynchronously.

In the more general case in which there are c coprocessors, numbered from 1 to c, the multiplication is expressed as $C = \alpha(AB_1 + \ldots + AB_{c+1}) + \beta(C_1 + \ldots + C_{c+1})$, and

$\alpha AB_i + \beta C_i$, with $1 \leq i \leq c$, is assigned to coprocessor i, and $\alpha AB_{c+1} + \beta C_{c+1}$ to the CPU.

Therefore, two important decisions need to be taken to obtain the lowest execution time: the distribution of the matrices in the computational system and the number of CPU threads to be generated. If the coprocessors are homogeneous, ideally the amount of data assigned to each coprocessor would be the same, but the overheads of coprocessor initialization and data transfer could make it preferable to use different sizes for different coprocessors. The second decision involves the number of CPU threads to generate, because the default option of using as many threads as available cores is not always the best one, for example, when the work assigned to the CPU is tiny and/or when the CPU has a high number of cores.

3 Auto-tuning method: empirical modeling

In this technique, a simple, but effective, model of the routine execution time is obtained and used to decide the values for a set of algorithmic parameters, AP, in order to minimize the total execution time [20]. In this case, the AP are of two types: the work distribution between the different computing units (the GPUs and the CPU, with the number of parameters equal to that of GPUs plus one) and the number of CPU threads.

For each GPU i, the computing time is modeled as:

$$T_{comp\_gpu\_i}(n, n_{gpu\_i}) = 2n^2 n_{gpu\_i} k_{comp\_gpu\_i} \tag{1}$$

where n is the dimension of the matrices to work with (A, B and C, considered square for simplicity), n_{gpu_i} determines the portion of matrix B assigned to GPU i (submatrix B_i has dimension $n \times n_{gpu\_i}$), and $k_{comp\_gpu\_i}$ is the computing system parameter, SP_{comp}, for the GPU i, defined as the average time to perform a basic operation (a double precision multiplication or addition) in this coprocessor.

The communication time for sending the operands from the CPU to GPU i and for receiving the results is modeled as:

$$T_{comm\_gpu\_i}(n, n_{gpu\_i}) = 3t_{s\_i} + (n^2 + 2nn_{gpu\_i})t_{w\_i} \tag{2}$$

where t_{s_i} and t_{w_i} (communication system parameters, SP_{com}) are the average times for starting the communication and for sending/receiving each double precision number to/from the GPU i.

Therefore, the total time to carry out the work assigned to GPU i is modeled as:

$$T_{gpu\_i}(n, n_{gpu\_i}) = T_{comm\_i}(n, n_{gpu\_i}) + T_{comp\_i}(n, n_{gpu\_i}) \tag{3}$$

In the same way, the CPU time, using t threads, would be:

 ISBN: 978-84-608-6082-2

$$T_{cpu}(n, n_{cpu}, t) = T_{comp\_cpu}(n, n_{cpu}) = 2n^2 n_{cpu} k_{comp\_cpu\_t} \quad (4)$$

where n_{cpu} determines the portion of matrix B assigned to the CPU. The computing system parameter $k_{comp\_cpu\_t}$ is the average time to perform a basic operation (a double precision multiplication or addition) in the CPU when t threads are used.

Given a specific problem to be solved, whose dimension is $n_r \times n_r$, and taking into account that the communications with the GPUs overlap with the CPU computing process, the goal of the auto-tuning method is to determine a set of values for the AP ($n_{gpu\_1}, ..., n_{gpu\_c}, n_{cpu}$ and t, with $n_r = n_{gpu\_1} + n_{gpu\_2} + ... + n_{cpu}$) that minimizes the total execution time, where this total time corresponds to that of the computing unit with the largest time to perform its assigned work:

$$\begin{aligned} &T_{total}(n_r, n_{gpu\_1}, ..., n_{gpu\_c}, n_{cpu}, t) = \\ &\max\{T_{gpu\_1}(n, n_{gpu\_1}), ...T_{gpu\_c}(n, n_{gpu\_c}), T_{cpu}(n, n_{cpu}, t)\} \end{aligned} \quad (5)$$

In general, our approach considers that the SP values depend on the quantity of data to work with/communicate. Therefore, at installation time, these are calculated on each computing unit (the CPU and each GPU) for a set of data sizes (*Unit_Problem_Size_Installation_Set*), using their corresponding benchmarks. After that, the second part of the installation consists of a complete search for the best AP values for each problem size in an established set of data sizes for the whole problem (*Global_Problem_Size_Installation_Set*). The results of the process make up the *Global_Best_AP_Installation_Set*.

At runtime, given a specific problem to be solved, of size $n_r \times n_r$, the AP values to use in its resolution are selected from the problem sizes in *Global_Best_AP_Installation_Set* closest to n_r.

4 Experimental results

A set of experiments varying the number and the kind of the GPUs was carried out in the 12CtwoC2075fourGTX590 platform. This is a shared-memory system with two hexacores (12 cores) Intel Xeon E5-2620, two GPU devices Nvidia Fermi Tesla C2075 with 5375 MBytes of Global Memory and 448 cores (14 Streaming Multiprocessors and 32 Streaming Processors per Multiprocessor), and four Nvidia GeForce GTX 590 with 1536 MBytes in Global Memory and 512 CUDA cores (16 Streaming Multiprocessors and 32 Streaming Processors per Multiprocessor). The different system configurations are referenced as 12CxC2075yGTX590, to indicate a CPU with 12 cores, x Fermi Tesla C2075 and y GeForce GTX 590.

As a guarantee of the model, Figure 1 shows a comparison in 12ConeC2075twoGTX590 of the execution time predicted by the model and the experimental execution time, for different problem sizes and for 10 different work distributions. The number of CPU threads was fixed to 10 for this set of experiments. The model offers good approximation of the execution times in almost all the cases, independently of the work distribution, and mainly for the biggest problem sizes. Therefore, it will be a useful tool to take quick decisions to reduce the total execution time.

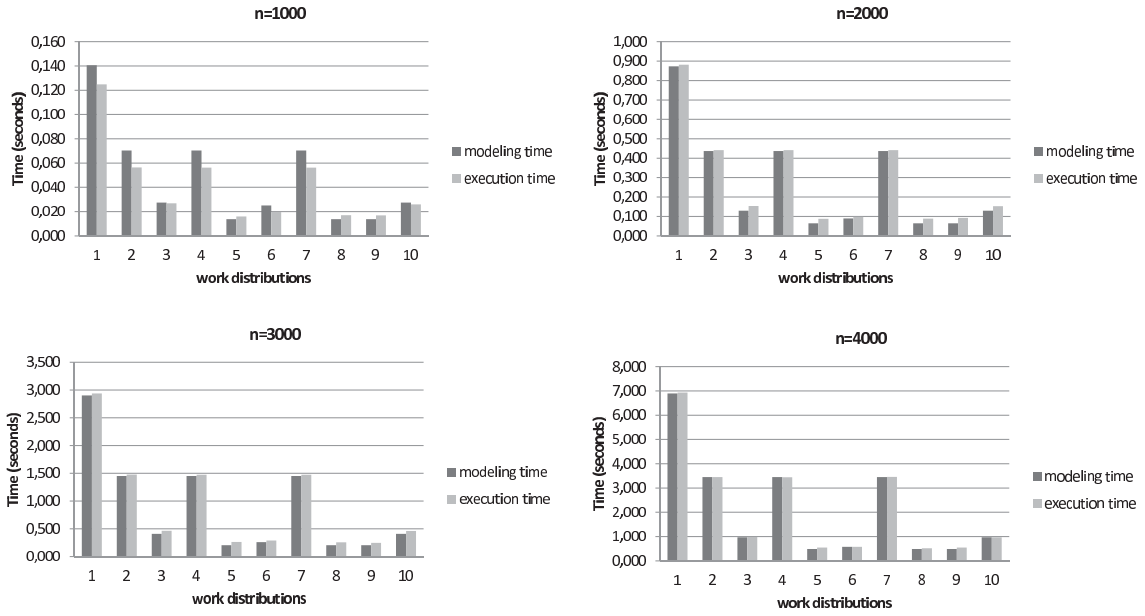


Figure 1: Comparison of model and execution times for different problem sizes and work distributions in 12ConeC2075twoGTX590

In order to measure how far the optimal solutions are from those achieved with the use of the auto-tuning methodology, another set of experiments was carried out. Due to the large quantity of different executions with all the possible combinations of AP values in the general approach, it was necessary to reduce the problem complexity for these experiments. So, the number of CPU threads was set to 10 and the problem size distribution grain between the computing units was established at 10% of the problem size. As shown in Table 1, using the empirical modeling technique, the obtained performance is very close to the optimum, mainly for large problem sizes, where the use of an optimization technique is more interesting.

Different work distribution strategies can be considered. Some, from the simplest to the smartest, are:

Table 1: Comparison of the value of the algorithmic parameters and the highest GFLOPS achieved for the hybrid `dgemm` routine. In 12CtwoC2075fourGTX590. The work distribution is shown as a percentage of the total size

	Optimum								Auto-Tuning							
n	n_1	n_2	n_3	n_4	n_5	n_6	n_{cpu}	GFLOPS	n_1	n_2	n_3	n_4	n_5	n_6	n_{cpu}	GFLOPS
1000	30	0	0	20	20	0	30	150	30	40	30	0	0	0	0	117
2000	10	20	10	10	20	20	10	312	20	20	20	20	20	0	0	264
4000	10	30	10	10	10	30	0	546	20	20	20	20	20	0	0	479
8000	10	30	0	10	10	30	10	652	10	30	10	10	10	30	0	642

- seqCPU: the routine is executed sequentially in the CPU.
- parCPU: the routine is executed in the CPU generating as many threads as available cores.
- bestGPU: the routine is executed in the fastest GPU of the system. The CPU is idle.
- allGPU: the routine is executed using all the GPUs of the system. The workload is distributed among them equally. The CPU is idle.
- Auto-Tuning: the routine is executed using the CPU and all the GPUs of the system. The distribution of the work and the number of CPU threads are decided automatically with the proposed methodology.

Table 2 shows a comparison of these strategies. The problem size ranges were established in accordance with the computing capacity of the different system configurations. As expected, the performance increases when the work distribution scheme is closer to the system architecture scheme, that is, with CPU parallelism to take advantage of the CPU cores and/or using the GPU capacities. In this way, scattering the work among all the GPUs is a very good solution. Finally, if the GPUs are combined with the CPU, and the distribution of the work and the selection of the number of CPU threads are performed automatically by the proposed method, the performance is increased even more in almost all the cases. It is important to note that this final proposed optimization is achieved without any code modification in the original routine, but only by adding the auto-tuning wrapper. Therefore, this approach seems to be a promising method for this kind of hybrid systems.

5 Conclusion and future work

This work studies the adaptation to multicore+multiGPU systems of the auto-tuning methodology called empirical modeling. It is applied to obtain balanced distributions of the

Table 2: Comparison of the GFLOPS achieved for the hybrid `dgemm` routine, using different work distribution methods. On different platform configurations

12ConeC2075oneGTX590										
	seqCPU	parCPU	bestGPU	allGPUs		Auto-Tuning				
n	GFLOPS	GFLOPS	GFLOPS	GFLOPS	GFLOPS	n_1	n_2	n_{cpu}	t	
1500	18	59	147	156	104	465	795	240	1	
2500	18	65	194	205	261	850	1325	325	10	
3500	19	72	201	220	269	1085	1155	1260	10	
4500	19	125	203	249	272	1260	1980	1260	10	
5500	19	141	226	247	377	1485	2530	1485	10	
6500	19	134	230	260	441	1820	2925	1755	8	

12ConeC2075twoGTX590										
	seqCPU	parCPU	bestGPU	allGPUs		Auto-Tuning				
n	GFLOPS	GFLOPS	GFLOPS	GFLOPS	GFLOPS	n_1	n_2	n_3	n_{cpu}	t
2500	18	60	198	299	165	600	1050	600	250	1
3500	18	70	214	319	377	875	1400	875	350,	10
4500	19	78	224	347	432	1080	1935	1080	405	10
5500	18	88	228	351	451	1265	1705	1265	1265	10
6500	19	87	240	379	514	1430	2340	1430	1300	10
7500	19	156	249	377	507	1650	2700	1650	1500	10

12ConeC2075threeGTX590											
	seqCPU	parCPU	bestGPU	allGPUs		Auto-Tuning					
n	GFLOPS	GFLOPS	GFLOPS	GFLOPS	GFLOPS	n_1	n_2	n_3	n_4	n_{cpu}	t
3500	18	70	192	396	400	770	1050	770	770	140	1
4500	19	75	206	419	474	900	1440	900	900	360	10
5500	19	94	236	448	499	1100	1925	1100	1100	275	10
6500	19	122	238	480	503	1300	1300	1300	1300	1300	10
7500	19	98	252	479	576	1350	2100	1350	1350	1350	10
8500	19	99	262	405	634	1530	2465	1530	1530	1445	10

12CtwoC2075fourGTX590													
	seqCPU	parCPU	bestGPU	allGPUs		Auto-Tuning							
n	GFLOPS	GFLOPS	GFLOPS	GFLOPS	GFLOPS	n_1	n_2	n_3	n_4	n_5	n_6	n_{cpu}	t
6500	19	120	235	655	715	845	1430	845	845	845	1430	260	10
7500	19	139	241	685	759	975	1800	975	975	975	1800	0	0
8500	19	115	244	670	780	1020	2040	1020	1020	1020	2040	340	10
9500	19	115	241	686	789	1330	2090	1330	1330	1330	2090	0	0
10500	19	99	248	718	802	1260	2100	1260	1260	1260	2100	1260	10
11500	19	124	251	699	818	1380	2300	1380	1380	1380	2300	1380	10

work and the best number of CPU threads to execute linear algebra routines. The methodology is explained through its application to a matrix-matrix multiplication. Satisfactory results for these complex, heterogeneous systems are reported. More experiments in more systems with larger numbers of coprocessors of different architectures and with other linear algebra routines are needed. More complex systems should be considered, for example multicore+multiGPU+multiMIC and heterogeneous clusters with heterogeneous nodes.

Acknowledgements

This work was supported by the Spanish MINECO, as well as by European Commission FEDER funds, under grant TIN2015-66972-C5-3-R.

References

[1] J. Bilmes, K. Asanovic, C. W. Chin, J. Demmel, Optimizing Matrix Multiply using PHIPAC: a Portable, High- Performance, ANSI C Coding Methodology, in: Proceedings of the International Conference on Supercomputing, ACM, 1997, pp. 340–347.

[2] J. Cuenca Muñoz, Optimización automática de software paralelo de álgebra lineal, (in Spanish), Ph.D. thesis, Departamento de Ingeniería y Tecnología de los Computadores de la Universidad de Murcia (2005).

[3] S. Hunold, T. Rauber, Automatic tuning of PDGEMM towards optimal performance, in: 11th International Euro-Par Conference, Lecture Notes in Computer Science, Vol. 3648, 2005, pp. 837–846.

[4] T. Katagiri, K. Kise, H. Honda, T. Yuba, Fiber: A generalized framework for auto-tuning software, Springer LNCS 2858 (2003) 146–159.

[5] R. C. Whaley, A. Petitet, J. Dongarra, Automated empirical optimizations of software and the ATLAS project, Parallel Computing 27 (1-2) (2001) 3–35.

[6] P. Alonso, R. Reddy, A. L. Lastovetsky, Experimental study of six different implementations of parallel matrix multiplication on heterogeneous computational clusters of multicore processors, in: M. Danelutto, J. Bourgeois, T. Gross (Eds.), PDP, IEEE Computer Society, 2010, pp. 263–270.

[7] Z. Chen, J. Dongarra, P. Luszczek, K. Roche, Self Adapting Software for Numerical Linear Algebra and LAPACK for Clusters, Parallel Computing 29 (2003) 1723–1743.

[8] J. Cuenca, L. P. García, D. Giménez, J. Dongarra, Processes distribution of homogeneous parallel linear algebra routines on heterogeneous clusters, in: Proc. IEEE Int. Conf. on Cluster Computing, IEEE Computer Society, 2005.

[9] J. Cuenca, D. Giménez, J. González, J. Dongarra, K. Roche, Automatic optimisation of parallel linear algebra routines in systems with variable load, in: PDP, IEEE Computer Society, 2003, pp. 409–416.

[10] Intel MKL web page, `http://software.intel.com/en-us/intel-mkl/`.

[11] IBM ESSL web page, `http://www-03.ibm.com/systems/software/essl/index.html`.

[12] K. Goto, R. A. van de Geijn, Anatomy of high-performance matrix multiplication, ACM Trans. Math. Softw. 34 (3).

[13] CULA GPU Accelerated Linear Algebra, `http://www.culatools.com/dense/performance/`.

[14] NVIDIA Corporation, NVIDIA CUDA CUBLAS Library Version 1.1, `http://docs.nvidia.com/cuda/cublas/`.

[15] E. Agullo, J. Demmel, J. Dongarra, B. Hadri, J. Kurzak, J. Langou, H. Ltaief, P. Luszczek, S. Tomov, Numerical linear algebra on emerging architectures: The PLASMA and MAGMA projects, Journal of Physics: Conference Series 180 (1).

[16] J. Cámara, J. Cuenca, L. P. García, D. Giménez, A. M. Vidal, Empirical installation of linear algebra shared-memory subroutines for auto-tuning, Journal of Parallel Programming 48 (3) (2014) 408–434.

[17] J. Cámara, J. Cuenca, L. P. García, D. Giménez, Auto-tuned nested parallelism: a way to reduce the execution time of scientific software in numa systems, Parallel Computing 40 (7) (2014) 309–327.

[18] L. P. García, J. Cuenca, D. Giménez, On optimization techniques for the matrix multiplication on hybrid CPU+GPU platforms, *Annals of Multicore and GPU Programming* 1 (1) (2014) 1–8.

[19] G. Bernabé, J. Cuenca, L. P. García, D. Giménez, Tuning basic linear algebra routines for hybrid CPU+GPUs platforms, in: *ICCS*, Procedia Computer Science, Vol. 29, 2014, pp. 30–39.

[20] G. Bernabé, J. Cuenca, L. P. García, D. Giménez, Auto-tuning techniques for linear algebra routines on hybrid platforms, Journal of Computational Science, Vol. 10, 2015, pp. 299-310.

[21] L. P. García, J. Cuenca, D. Giménez, F.J. Herrera, On guided installation of basic linear algebra routine in nodes with manycore components, in: Proceedings of 7th International Workshop on Programming Models and Applications for Multicores and Manycores (*PMAM*), 2016.

[22] M. Fatica, Accelerating Linpack with CUDA on heterogenous clusters, in: Proceedings of 2nd Workshop on General Purpose Processing on Graphics Processing Units, GPGPU-2, ACM, New York, NY, USA, 2009, pp. 46–51.

[23] S. Ohshima, K. Kise, T. Katagiri, T. Yuba, Parallel processing of matrix multiplication in a CPU and GPU heterogeneous environment, in: Proceedings of the 7th International Conference on High Performance Computing for Computational Science, VECPAR'06, Springer-Verlag, 2007, pp. 305–318.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Fast Intra Mode Decision for an H.264/AVC to HEVC Video Transcoder

Antonio Jesus Diaz-Honrubia[1], Jose Luis Martinez[1] and Pedro Cuenca[1]

[1] *Albacete Research Institute of Informatics (I3A), University of Castilla-La Mancha, Albacete, Campus Universitario s/n, Spain.*

emails: Antonio.DHonrubia@uclm.es, JoseLuis.Martinez@uclm.es, Pedro.Cuenca@uclm.es

Abstract

The H.264/*Advanced Video Coding* (AVC) standard has been widely used in the last years, leading to a lot of video streams which are currently encoded in this format. However, the new *High Efficiency Video Coding* (HEVC) standard was developed by the *Joint Collaborative Team on Video Coding* to replace it and it is thought to dominate the market for the next few years. In terms of rate-distortion (RD) performance, HEVC roughly doubles the RD compression performance of H.264/AVC at the expense of a high computational cost. These facts make efficient transcoding algorithms between H.264/AVC and HEVC necessary. Moreover, intra encoded streams are commonly used in certain scenarios, such as video editing and post-production, making a migration of these intra contents from H.264/AVC to HEVC necessary. This paper proposes a fast intra H.264/AVC to HEVC transcoding algorithm based on intra mode detection based on the decisions made by H.264/AVC. Experimental results show that the proposed algorithm achieves a good trade-off between coding efficiency and complexity compared with the anchor transcoder, and, moreover, it outperforms non-transcoding algorithms.

Key words: HEVC, H.264/AVC, Transcoding, Intra Prediction.

1 Introduction

In the last ten years, the H.264/*Advanced Video Coding* (AVC) standard [1] has been the most extended video compression standard for High Definition (HD) video coding and, thus, there are a lot of contents according to this standard. However, in January 2013, the *High Efficiency Video Coding* (HEVC) standard [2] was established by the *Joint Collaborative*

Team on Video Coding (JCT-VC) as a natural evolution of H.264/AVC. HEVC was conceived with the purpose of achieving a highly efficient performance for delivering high quality multimedia services over bandwidth-constrained networks, but also to give support to *Ultra High Definition* (UHD), which demand a high bandwidth. In terms of Rate-Distortion (RD) performance, HEVC roughly doubles the RD compression performance of H.264/AVC but at a cost of extremely high computational and storage complexities during encoding [3].

Furthermore, there are a wide range of kind of contents that can be encoded, such as films, video conferencing, streaming or recorded screen sequences. Specifically, films are usually encoded under different configurations depending on their usage. In versions which are oriented to the consumer, a good *Group Of Pictures* (GOP) pattern might be an intra frame followed by several inter frames in order to provide random access to the video stream every few seconds. In other scenarios, such as video editing or post-processing, it is necessary to access each frame separately, without the need to decode adjacent frames. In this scenario a GOP pattern which is only composed of intra frames is needed.

Considering both the superior compression performance of HEVC, as well as the large body of content that is currently encoded using the H.264/AVC standard, a transcoder that can convert H.264/AVC bitstreams into HEVC bitstreams is interesting device for efficient conversion between the H.264/AVC standard and the HEVC. As it will be shown in Section 2, there are several proposals which aims at accelerating an inter H.264/AVC to HEVC transcoder, however, there are not many proposals for the intra transcoder, which is also useful.

With this fact in mind, this paper presents an intra H.264/AVC to HEVC transcoder. Specifically, the proposals accelerate the intra mode decision module, which has not been tackled before in the transcoding scenario. The algorithm re-uses the information collected in the H.264/AVC decoding process in order to detect the HEVC intra mode direction.

The experimental results show that the proposed algorithm, when compared with the anchor intra transcoder, achieves a time reduction of almost 20% on average, with a negligible 0.8% loss in efficiency in terms of BD-rate [4], which measures the increment of bit rate while preserving the same objective quality of the video stream.

The remainder of this paper is organized as follows. Section 2 includes some technical background to the new HEVC standard focusing on intra coding, and the related work which has been previously carried out. Section 3 presents the proposed algorithm, and then the experimental results are given in Section 4. Finally, Section 5 concludes the paper.

2 Technical Background and Related Work

HEVC maintains the block-based hybrid approach used in H.264/AVC and previous video compression standards. In addition, new tools have been introduced in HEVC that increase its coding efficiency compared with it. One of the most important changes affects picture

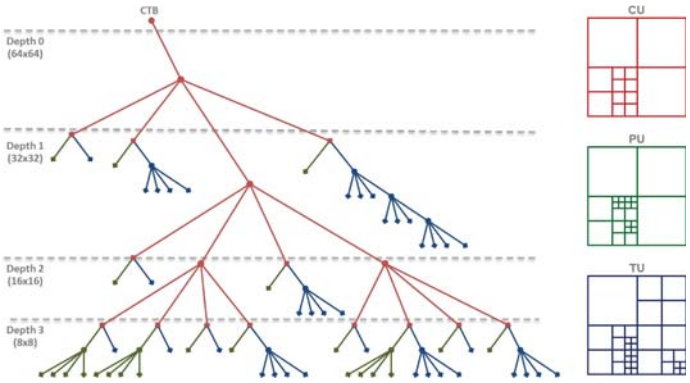


Figure 1: Partitioning of CTU into CUs, PUs and TUs.

partitioning. HEVC defines a new flexible *Coding Tree Unit* (CTU) structure with the aim of achieving an optimal adaptation to the content details. The CTU's maximum size is 64x64 pixels and it may contain one or more *Coding Units* (CUs), which are a replacement for the 16x16 pixel *MacroBlocks* (MBs) used in the previous standards. A CU size can vary from 64x64 ($depth = 0$) to 8x8 ($depth = 3$) pixels, as it can be iteratively partitioned into four squared sub-CUs. Therefore, a CTU can be partitioned into four *Depth Levels*, from $d = 0$ for 64x64 CUs to $d = 3$ for 8x8 CUs. Thus, a CU in depth level d can be denoted as $CU_{d,k}$ ($k = 0, 1, ..., 4^d - 1$), and the four sub-CUs pending on $CU_{d,k}$ are denoted as $CU_{d+1,4k+i}$ (i=0,1,.., 3).

Thus, each $CU_{d,k}$ becomes a root of two new trees which contain two new unit types: the *Prediction Units* (PUs), and the *Transform Units* (TUs). For inter prediction, 8 different PU sizes are checked, while in the case of intra prediction, a PU uses the same 2Nx2N size as for the $CU_{d,k}$ to which it belongs, allowing it to be split into four NxN PUs only for CUs at the deepest level. Therefore, the PU size can range from 64x64 to 4x4 pixels. After that, the prediction residue is transformed using various TU sizes from 32x32 to 4x4.

In Figure 1, an example of the partitioning is shown, depicting how a CTU is structured in a hierarchical tree where each CU branch ends in a leaf ($CU_{d,k}$) that is the root for the two new prediction and transform trees, containing the PU and TU trees.

It should be noted that a CTU can be split into 149 different intra PUs ($\sum_{d=0}^{4} 4^d$), and each of these PUs should be evaluated for 33 directional modes and 2 non-directional modes (*DC* and *Planar*). The computational complexity difference with respect to H.264/AVC can be noted, since it only allowed up to 9 modes and the *DC* mode (*Planar* was also available in 16x16 intra blocks, but in this case only 2 directional modes were checked).

However, regarding the selection, the HEVC reference software, namely HM [5], implements a fast mode decision based on *Piao et al.*'s scheme [6], which performs a fast *pseudo Rate-Distortion Optimization* (RDO), the so-called *Rough Mode Decision* (RMD), over all

modes and then it only performs the exhaustive RDO on 3 of them when the PU size is from 64x64 to 16x16 pixels, and on 8 of them when the PU is smaller.

The above analysis highlights the need to reduce the complexity for HEVC intra prediction in an H.264/HEVC video transcoder. With the algorithm approach, proposals can be classified into two categories. Algorithms in the first one reduce the number of CU sizes that need to be checked, mainly by limiting the depth of the CTU tree, and thus they can be named as tree pruning algorithms. The second category reduces the number of angular prediction modes to be checked in the RMD and RDO stages.

The intra H.264/AVC to HEVC transcoding problem has not been addressed in depth in previous works. In the first category of proposals, in the specific framework of an intra transcoder, the authors use *Support Vector Machine* classifiers to assist decisions on the quadtree splitting. Specifically, they use a vector which includes the number of bits that H.264/AVC used to encode the sub-CUs of the current CU, the average and the variance of the residue of the three most dominant directional modes and the QP. Then, they obtain confidence levels of how accurate it would be not to further split the CU, testing the proposal with different accuracies, obtaining a 30% HEVC encoding time reduction.

A more recent work, in [7] the authors present an intra transcoding algorithm. In that paper, a quadtree pruning algorithm based on bayesian classifiers is presented. Some features are fetched in the H.264/AVC decoding stage and, then, it is re-used as input for Bayesian classifiers which have been previously built. This intra transcoder achieves a 57% HEVC encoding time reduction with a BD-rate of 2.2%.

Regarding the second category of proposals, the authors could not find other works at the time of writing which aims to accelerate an intra H.264/AVC to HEVC transcoder. One of the most recent proposals (not in a transcoding scenario but in a fast HEVC encoding one) is based on obtaining the dominant gradient of the texture in order to try only a few directional intra modes [8]. Specifically, the authors define 12 orientations and choose one of them on the basis of the content of the PU. They then only perform the RDO on 2 to 4 directional modes (those modes whose orientation is the closest to the orientation of the gradient).

Finally, again in a fast intra HEVC encoding scenario, [9] proposes a combination of both kinds of algorithms. The authors suggest a fast-partitioning decision algorithm that uses the correlation of the content and the optimal CTU tree depth, limiting the minimum and maximum depth levels. The algorithm is based on the observed evidence that small CUs tend to be chosen for rich textured regions and vice versa. It computes a depth predictor by using the neighboring tree blocks and 2 early terminations for prediction modes based on the statistics of neighboring blocks and the RDO cost of the candidates. The authors report a time reduction of 21% with a BD-rate of 1.7%.

Table 1: Number of RDM and RDO modes checked by the *Piao et al.*'s and the proposed algorithms.

PU size	#(RMD directional modes)		#(RDO directional modes)	
	Piao et al. [6]	Proposed	*Piao et al.* [6]	Proposed
64x64	33	5	3	2
32x32	33	5	3	2
16x16	33	7	3	2
8x8	33	11	7	5
4x4	33	11	7	5

3 Proposed Intra Transcoding Algorithm for Direction Detection

The proposed algorithm follows the same scheme as *Piao et al.*'s algorithm [6] which is already implemented in HM reference software, as commented in Section 2. Nevertheless, while the original *Piao et al.*'s algorithm checks all the possible directional modes in the RMD (fast) stage, the proposed algorithm only tries a subset of them. Furthermore, the proposed algorithm also decrements the number of modes in which the full RDO is checked. Table 1 shows the number of modes checked at the RMD and RDO stages in both algorithms according to the PU size. It can be seen that in both cases the number of RDM and RDO to be checked increases for smaller PU sizes, since the smaller the PU is, the more difficult is to make a good prediction. Furthermore, in both algorithms the 2 non-angular modes (i.e. DC and Planar modes) are always checked in the RDM stage.

First of all, it must be taken into account that the HEVC and H.264 partitions sizes are generally not the same (it might be the same when the PU size is 16x16, 8x8 or 4x4, but even in these cases, this need not be true). Because of this reason a mapping between the HEVC PUs and H.264/AVC MBs is performed, so that a PU contains one or more MBs. If the PU size is bigger than the MB size, the PU will contain several MBs, while if the size is the same (i.e., 16x16 pixels), the correspondence is one to one. Finally, if the PU size is smaller than an MB, then the only overlapping MB is considered as the only sample for the current PU.

When assigning a PU to the different overlapping H.264/AVC directional modes, the original numeration in H.264/AVC is changed so that the new number represents the HEVC direction whose orientation is the closest to the orientation of the H.264/AVC mode (DC and planar modes are not taken into account for calculations). Figure 2 shows the correspondence between the HEVC and H.264/AVC numerations.

Then, if more than one MB overlaps the current PU (i.e., the PU size is 64x64 or 32x32 pixels), the average and the variance of all the overlapping modes are calculated and the average (rounded to integer) is used as a directional predictor in HEVC. For these two PU

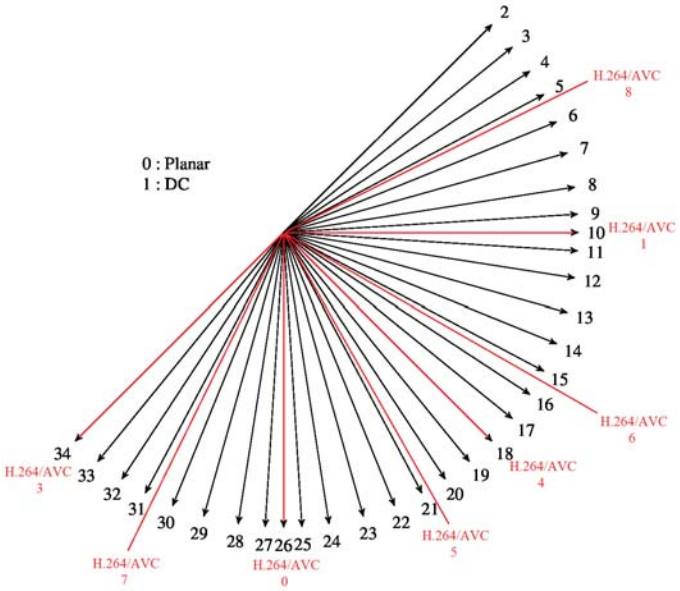


Figure 2: Correspondence between HEVC and H.264/AVC intra directional modes.

sizes, the intra 16x16 MBs are not considered, since these MBs can only choose between 2 directional modes, either horizontal or vertical (in smaller PUs they are used as they are the only available sample). In this case, if a PU does not contain any sample (i.e., all the MBs were 16x16 or not angular modes), the proposed algorithm is not used and the default *Piao et al.*'s algorithm is used.

If the PU size was 16x16 or 8x8 (there is only one MB which overlaps the current PU), the directional predictor is the mode of the only available sample. In the case of a 4x4 PU, as the area of the image is too small and the information of the only available MB might not be very accurate, the average and the variance of the current and all the adjacent PUs (up to 8 PUs) are calculated so that a global gradient is taken into account, and the average (rounded to integer) is used as a directional predictor.

If the variance calculated (for the PU sizes where a variance is calculated) is higher than a threshold[1], then the current PU is considered as a high entropy PU and the default [6] fast HM algorithm is used. Otherwise, the selected predictor and $(n-1)/2$ modes to its left and other $(n-1)/2$ modes to its right are checked with the RMD, where n represents the number of modes used to be checked in the RMD stage, as described in Table 1. Then, the best m modes in the RMD stage, where m is the number of modes used to check the RDO, as described in Table 1. DC and Planar modes are always checked in the RMD since

[1]A threshold of 120 has been chosen in the experiments that we performed.

it is very common to choose one of them, and predicting them is not easy, something which would result in errors.

4 Performance Evaluation

This section aims to evaluate the intra transcoding algorithm presented in this paper. In order to ensure a common framework for the simulations, the JCT-VC defined a set of common test conditions [10] to homogenize comparisons between experiments. Therefore, this performance evaluation has been carried out in accordance with these guidelines. Specifically, the QP values used were $\{22, 27, 32, 37\}$ with the *All Intra* (AI) configuration. The sequences contained in this document are grouped into *Classes*, according to their resolution:

- Class A (2560x1600 pixels): *Traffic* and *PeopleOnStreet*.
- Class B (1920x1800 pixels): *Kimono*, *ParkScene*, *Cactus*, *BQTerrace* and *BasketballDrive*.
- Class C (832x480 pixels): *RaceHorsesC*, *BQMall*, *PartyScene* and *BasketballDrill*.
- Class D (416x240 pixels): *RaceHorses*, *BQSquare*, *BlowingBubbles* and *BasketballPass*.
- Class E (1280x720 pixels): *FourPeople*, *Johnny* and *KristenAndSara*.

The results are presented in terms of *time reduction* and *BD-rate* [4]. The global *Peak Signal-to-Noise Ratio* ($PSNR_{YUV}$) for the BD-rate was calculated as the weighted average of the luma PSNR ($PSNR_Y$) and chroma PSNRs ($PSNR_U$, $PSNR_V$), according to Equation 1 (weights recommended by the JCT-VC in [11], which are considered a fair representation of the human visual quality).

The software used was JM 18.4 [12] for H.264/AVC, a modified version of the HM 16.2 [5] encoder including the proposed mode detection algorithm, and the original HM 16.2 (used as anchor to calculate the BD-rate and the time reduction). The remainder of the coding parameters not mentioned are kept as default in the configuration file. All the measurements were performed on a six-core Intel Core i7-3930K CPU running at 3.20GHz.

$$PSNR_{YUV} = \frac{6 \times PSNR_Y + PSNR_U + PSNR_V}{8} \tag{1}$$

Table 2 contains the results of proposed algorithm using the AI configuration. It can be seen that it achieves a very low BD-rate with a time reduction of around 20%. It should be noticed that the time reduction is not very high since it is compared with the already fast algorithm implemented in HM [6], so it represents a 20% acceleration of an already fast algorithm, which is not a bad value.

Regarding the variability of the results, the time reduction is very similar in all the cases, ranging from 17% to 22%, while the variability of the BD-rate is higher since some

Table 2: Time reduction and coding efficiency results of the proposed algorithm.

		BD-rate (%)	Time reduction (%)
Class A	Traffic	0.8	21.00
	PeopleOnStreet	1.1	17.08
	Nebuta	0.1	17.40
	SteamLocomotive	0.3	20.31
Class B	Kimono	0.5	20.80
	ParkScene	0.6	20.20
	Cactus	1.1	20.25
	BasketballDrive	1.1	22.23
	BQTerrace	0.8	19.89
Class C	BasketballDrill	1.0	18.65
	BQMall	1.2	19.92
	PartyScene	1.2	17.86
	RaceHorsesC	0.7	18.72
Class D	BasketballPass	1.2	18.09
	BQSquare	1.2	16.48
	BlowingBubbles	1.0	16.87
	RaceHorses	0.8	17.04
Class E	FourPeople	1.0	20.87
	Johnny	1.0	19.50
	KristenAndSara	1.0	19.50
Arithmetic mean		0.9	19.13
Standard deviation		0.3	1.60
Coefficient of variation (%)		33.3	8.36

sequences obtain very low BD-rate. This fact can be statistically demonstrated with the *Coefficient of Variation* (CV) which, as shown in (2), measures the standard deviation (σ) as a percentage of the arithmetic mean ($\bar{x}$). Even though the standard deviation also provides a dispersion metric, it is dependent of the scale of the variable, while the CV is independent of the scale and can be used to compare the dispersion of different variables, such as the BD-rate and the time reduction. Thus, while the CV of time reduction is only 8%, the CV of the BD-rate is 33%. This fact is a good point of the proposed algorithm, since it means that for some sequences (e.g. *Nebuta*, *SteamLocomotive*, ...), while the time reduction is still close to the average value, the BD-rate obtains much better results.

$$CV(\%) = \frac{\sigma}{\bar{x}} \times 100 \tag{2}$$

4.1 Comparison with a non-transcoding algorithm

Regarding the comparison with other works, as said in Section 2, the authors could not find other works which try to accelerate the intra direction decision in the transcoding scenario. Because of this, the algorithm proposed by *Ruiz et al.* [8] has been executed in

Table 3: Comparison of the proposed algorithm with *Ruiz et al.*'s algorithm [8].

	Proposed algorithm		*Ruiz et al.*'s algorithm [8]	
	BD-rate (%)	Time reduction (%)	BD-rate (%)	Time reduction (%)
Traffic	0.8	21.00	1.3	16.69
Kimono	0.5	20.80	0.6	15.71
BasketballDrill	1.0	18.65	2.1	14.16
BlowingBubbles	1.0	16.87	1.6	12.72
FourPeople	1.0	20.87	1.2	17.21
Average	0.9	19.64	1.4	15.30

the transcoding scenario. Table 3 shows the results of both algorithms for some sequences (one for each class). First of all, it can be noted that the BD-rate in this scenario is much higher than that reported in [8]. This is due to the fact that the sequences are subjected to two quantification processes. Furthermore, it can be seen that the proposed intra direction detection algorithm obtains 4% more in time reduction with 0.5% less of losses in BD-rate terms, what demonstrates that the information gathered in the H.264/AVC decoding stage of the transcoder improves the performance of the direction detection for intra PUs.

5 Conclusions and Future Work

In this paper we have presented an intra H.264/AVC to HEVC transcoding algorithm which tries to further accelerate the already accelerated decision on the intra directional mode to be chosen using the intra modes which were decided in H.264/AVC. The proposed algorithm can reduce the computational complexity of the intra H.264/AVC to HEVC transcoder by 20%, with only a slight BD-rate increase of 0.8%. We would like to highlight the novelty of the scenario in view of the lack of consistent proposals for such an intra transcoder.

As future work in the topic, it can be noticed that the proposed algorithm can be combined with an early CU termination algorithm, such as those described in Section 2. It is clear that, even though both algorithms do not have the same target, they would not be fully independent of each other, since the proposed algorithm would not be applied to those branches of the quadtree pruned by early termination algorithm. Thus, neither the resulting BD-rate nor the time reduction would be the sum of those of the two algorithms.

Acknowledgements

This work has been supported by the MINECO and European Commission (FEDER funds) under the project TIN2015-66972-C5-2-R and by the Spanish Education, Culture and Sports Ministry under grant FPU12/00994.

References

[1] ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC) version 22, "Advanced Video Coding for Generic Audiovisual Services," Feb 2012.

[2] B. Bross, W. Han, J.-R. Ohm, G. J. Sullivan, Y.-K. Wang, and T. Wiegand, "High Efficiency Video Coding (HEVC) Text Specification Draft 10," Doc. JCTVC-L1003, Jan 2013.

[3] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T.-K. Tan, and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, Dec 2012.

[4] G. Bjøntegaard, "Improvements of the BD-PSNR Model," *ITU-T SG16 Q6 Document VCEG-AI11, 35th VCEG Meeting*, Jul 2008.

[5] "HM Reference Software," https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/.

[6] Y. Piao, J. Min, and J. Chen, "Encoder Improvement of Unified Intra Prediction," in *Proc. 3th JCT-VC Meeting, Doc. JCTVC-C207*, October 2010.

[7] A. J. Diaz-Honrubia, J. L. Martinez, P. Cuenca, and H. Kalva, "A Fast Splitting Algorithm for an H.264/AVC to HEVC Intra Video Transcoder," in *Data Compression Conference (DCC 2016), Salt Lake City, USA*, Mar 2016.

[8] D. Ruiz, G. Fernandez-Escribano, J. Martinez, and P. Cuenca, "Fast Intra Mode Decision Algorithm Based on Texture Orientation Detection in HEVC," *Signal Processing: Image Communication*, vol. PP, 2016.

[9] P. A. L. Shen, Z. Zhang, "Fast CU Size Decision And Mode Decision Algorithm for HEVC Intra Coding," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 1, pp. 207–213, Feb 2013.

[10] F. Bossen, "Common HM Test Conditions and Software Reference Configurations," in *Proc. 12th JCT-VC Meeting, Doc. JCTVC-L1100*, Jan 2013.

[11] G. J. Sullivan and K. Minoo, "Objective quality metric and alternative methods for measuring coding efficiency ," in *Proc. 8th JCT-VC Meeting, San Jose, USA, JCTVC-H0012*, July 2013.

[12] "JM Reference Software, version 18.4," http://iphome.hhi.de/suehring/tml/.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Phase-fitted splitting methods for oscillatory genetic regulatory systems

Julius Osato Ehigie[1], Ruqiang Zhang[2], Xilin Hou[1] and Xiong You[2]

[1] *College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China*

[2] *College of Sciences, Nanjing Agricultural University, Nanjing 210095, China*

emails: jehigie@unilag.edu.ng, 2014111001@njau.edu.cn, hxl@njau.edu.cn, youx@njau.edu.cn

Abstract

Phase-fitted splitting methods are constructed for the numerical solution of oscillatory differential equations with an additively separable structure. The result of the experiment on a two-gene regulatory network shows that the new phase-fitted splitting methods (PLT, PStrang, PTJ) are more effective than the corresponding traditional splitting methods and Runge-Kutta methods of the same order.

Key words: Gene regulatory network, splitting methods, phase fitting
MSC 2000: 65L04, 65L05, 65L06

1 Introduction

Almost all the organisms have develop a mechanism to adapt to the daily sunlight rhythm. This mechanism consists of a molecular network, in particular, a gene network, with interacting components. In most cases, negative feedback loops in a gene network have been shown to entrain sustained oscillations which play an important role in the functioning of living cells. The dynamics of an N-gene regulated network can be modelled by a system of ordinary differential equations

$$\begin{aligned} \dot{r}(t) &= -\Gamma r(t) + F(p(t)), \\ \dot{p}(t) &= -Mp(t) + Kr(t), \end{aligned} \tag{1}$$

where $r(t)$ and $p(t)$ are N-dimensional vectors representing the concentrations of mRNAs and proteins at time t, respectively, $F(p(t))$ is the vector of regulatory functions determining

 ISBN: 978-84-608-6082-2

the cooperativity of two genes, Γ, M and K are diagonal matrices of reaction rate constants. To simulate the qualitative behavior of the system (1), biologists employ the standard Runge-Kutta (RK) methods which are available in most scientific computation softwares. However, the general purpose RK methods fail to take into account the special structure of the system (1) and can not produce satisfactory numerical results. Recently, regarding the additively separable structure of the system (1), You et. al. [1] took a splitting approach and obtained very accurate results. For details of splitting methods, see Blanes [2].

The purpose of this paper is to adapt the standard splitting methods with phase-fitting to the system (1) whose solutions share an oscillatory property.

2 Phase-fitted splitting methods for oscillatory systems

Consider the initial value problem (IVP) of ordinary differential equations

$$\dot{y} = f(y), \quad t > 0, \quad y(0) = y_0. \tag{2}$$

Suppose we can decompose the vector field f in (2) into two subfields

$$y' = f^{[1]}(y) + f^{[2]}(y). \tag{3}$$

and assume that each subfield $f^{[i]}(y),\ i = 1, 2$ can has the exact flow $\varphi_t^{[i]}$. The *Lie-Trotter splitting method*, which is of order 1, is defined by

$$\Psi_h^{[LT]} = \varphi_h^{[2]} \circ \varphi_h^{[1]}. \tag{4}$$

A composition of $\Psi_h^{[LT]}$ with its *adjoint* $\Psi_h^{[LT]*} := \Psi_{-h}^{[LT]^{-1}} = \varphi_h^{[1]} \circ \varphi_n^{[2]}$ gives the *Strang splitting method*

$$\Psi_h^{[ST]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}. \tag{5}$$

which is symmetric and has order two. Composition of the Strang splitting leads to the *triple jump*

$$\Psi_h^{[TJ]} = \Psi_{\gamma_3 h}^{ST} \circ \Psi_{\gamma_2 h}^{ST} \circ \Psi_{\gamma_1 h}^{ST} \tag{6}$$

where $\gamma_1 = \gamma_3 = \frac{1}{2-2^{1/3}}$ and $\gamma_3 = -\frac{2^{\frac{1}{3}}}{2-2^{\frac{1}{3}}}$. This method is also symmetric and has order 4.

Consider the scalar autonomous test system of the form

$$\dot{q} = \omega p, \quad \dot{p} = -\omega q \tag{7}$$

whose vector field can be decomposed as $f^{[1]} + f^{[2]}$ with

$$f^{[1]} = \begin{pmatrix} \omega p \\ 0 \end{pmatrix}, \quad f^{[2]} = \begin{pmatrix} 0 \\ -\omega q \end{pmatrix}.$$

Applying a splitting method yields

$$(q_{n+1}, p_{n+1})^T = R(\nu)(q_n, p_n)^T, \tag{8}$$

 ISBN: 978-84-608-6082-2

Definition 2.1 *The quantities*

$$PL(\nu) := \nu - \arccos \frac{\mathrm{tr}(R(\nu))}{2\sqrt{\det(R(\nu))}} \qquad D(\nu) := 1 - \sqrt{\det(R(\nu))} \tag{9}$$

are called the dispersion (or phase lag) and the dissipation (or amplification factor error) of the splitting or composition method, respectively. If $P(\nu) = 0$ and $D(\nu) = 0$, the method is zero-dispersive (or phase-fitted) and zero-dissipative (amplification-fitted), respectively.

It is easy to show that the Lie-Trotter method, Strang and triple jump methods are zero-dissipative. Phase-fitted splitting methods are listed as follows:

(i) The phase-fitted Lie-Trotter method (PLT) of order one

$$\Psi_h^{[PLT]} = \phi^{[2]}_{b(\nu)h} \circ \phi^{[1]}_{a(\nu)h} \tag{10}$$

with $a(\nu) = b(\nu) = \frac{2\sin(\nu/2)}{\nu}$.

(ii) The phase-fitted Strang splitting method (PST) of order two

$$\Psi_h^{[PST]} = \phi^{[1]}_{a(\nu)h/2} \circ \phi^{[2]}_{b(\nu)h} \circ \phi^{[1]}_{a(\nu)h/2} \tag{11}$$

with $a(\nu) = \frac{2\sin\nu}{\nu(\cos\nu+1)}$ and $b(\nu) = \frac{\sin\nu}{\nu}$.

(iii) The phase-fitted triple jump composition method (PTJ) of order four

$$\Psi_h^{[PTJ]} = \Psi^{ST}_{\gamma_3(\nu)h} \circ \Psi^{ST}_{\gamma_2(\nu)h} \circ \Psi^{ST}_{\gamma_1(\nu)h} \tag{12}$$

where

$$\begin{aligned}
\gamma_1(\nu) = \gamma_3(\nu) = {} & \tfrac{1}{2-2^{1/3}} - \left(\tfrac{1}{108} + \tfrac{1}{108}2^{2/3} + \tfrac{1}{54}2^{1/3}\right)\nu^2 + \left(\tfrac{5}{216} + \tfrac{1537}{90720}2^{2/3} + \tfrac{1067}{45360}2^{1/3}\right)\nu^4 + \mathcal{O}(\nu^6) \\
\gamma_2(\nu) = {} & -\tfrac{2^{1/3}}{2-2^{1/3}} + \left(\tfrac{1}{54} + \tfrac{1}{54}2^{2/3} + \tfrac{1}{27}2^{1/3}\right)\nu^2 - \left(\tfrac{31}{1080} + \tfrac{253}{11340}2^{2/3} + \tfrac{94}{2835}2^{1/3}\right)\nu^4 + \mathcal{O}(\nu^6)
\end{aligned}$$

3 Numerical simulation of the Goodwin oscillator

The new methods will be applied to the two-gene cross-regulatory network

$$\dot{r}_1 = -\gamma_1 r_i + m_i H_i(p_i), \quad \dot{p}_i = -\mu_i p_i + k_i r_i, \quad i = 1, 2 \tag{13}$$

with the Hill functions $H_i = \frac{1-(-1)^{3-i}}{2} + (-1)^{3-i}\frac{\theta_{3-i}^{n_{3-i}}}{\theta_{3-i}^{n_{3-i}} + p_{3-i}^{n_{3-i}}}$, $i = 1, 2$. The parameters are taken as $m_1 = 1.8$, $m_2 = 1.8$, $n_1 = 3$, $n_2 = 3$, $k_1 = 1$, $k_2 = 1$, $\gamma_1 = 1$, $\gamma_2 = 1$, $\mu_1 = 1$, $\mu_2 = 1$, $\theta_1 = 0.6542$, $\theta_2 = 0.6542$. We integrate the network on the time interval $[0, 100]$ with step sizes $h = 2^j, j = -7, -6, -5, -4$. Traditional Runge-Kutta methods the Euler

 ISBN: 978-84-608-6082-2

Table 1: The maximal global errors produced by the six methods for the two-gene regulatory network.

h	Euler	Heun	RK 4	PLT	PStrang	PTJ
1/2	8.1e-1	2.7e-1	1.8e-4	2.1e-2	7.1e-4	1.7e-4
1/4	5.2e-1	6.8e-2	7.1e-5	1.1e-2	2.9e-4	1.2e-5
1/8	1.9e-1	1.7e-2	3.5e-6	5.8e-3	5.1e-5	7.2e-8
1/16	6.6e-2	4.3e-4	2.0e-7	2.9e-3	1.3e-5	4.5e-8

method of order one, the Heun method of order two and the four-stage fourth-order Runge-Kutta method (denoted by RK4) are also utilized for comparison. The numerical result is presented in Table 1.

From Table 1 we can see that the new phase fitted splitting methods are more effective than the traditional Runge-Kutta methods of the same order. The first order PLT is more accurate that the first order Euler method and the second order Heun method, and the fourth order PTJ is more accurate than the four order Runge-Kutta method. Moreover, for comparatively large step sizes $h = 1/2, 1/4$, the second order PStrang performs almost as well as the fourth order Runge-Kutta method.

Acknowledgements

This research was partially supported by NSFC under Grant No. 11171155.

References

[1] X. You, X. Liu and I.H. Musa, *Splitting strategy for simulating genetic regulatory networks*. Comput Math Methods Med, **2014** (2014) Article ID 683235, 9 pages.

[2] S. Blanes, F. Casas and A. Murua, *Splitting and composition methods in the numerical integration of differential equations*, Bol. Soc. Esp. Mat. Apl. **45** (2008) 89–145.

[3] Z. Chen, X. You, X. Shu and M. Zhang, *A new family of phase-fitted and amplification-fitted Runge-Kutta type methods for oscillators*, *J. Appl. Math.* **2012** (2012) Article ID 236281, 27 pages.

[4] A. Polynikis, S. J. Hogan and M. di Bernardo, Comparing different ODE modelling approaches for gene regulatory networks, *J. Theor. Biol.* **261** (2009) 511–530.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A model for Leshmaniasias disease transmission considering asymptomatics and reservoirs

Lourdes Esteva[1], Cristobal Vargas[2] and Cruz Vargas de León[3]

[1] *Facultad de Ciencias, Universidad Nacional Autónoma de México*

[2] *Departamento de Control Automático, Centro de Investigación y de Estudios Avanzados-IPN*

[3] *Unidad de Medicina Experimental, Hospital General Dr. Eduardo Liceaga*

emails: lesteva@ciencias.unam.mx, cvargas@ctrl.cinvestav.mx, leoncruz82@yahoo.com.mx

Abstract

Leshmaniasis is a parasite disease transmitted by the bites of sand fly bites. Cutaneous Leshmaniasis is the most common form of the disease and it is endemic in the Americas. Around 70 animal species, including humans, have been found as natural reservoir hosts of Leshmania parasites. Among the reservoirs, dogs are the most important ones due to their proximity to the human habitat. Infection by leshmaniasis does not invariably cause illness in the host, and can remain asymptomatic for a long period specially in dogs.

In this work we formulate a model to study the transmission of the disease among the vector, humans and dogs. Our main objective is to asses the impact of asymptomatics and dogs on the spread of Leshmaniasis. For this end we calculate the Basic Reproduction Number of the disease and we carry out sensitivity analysis of this parameter with respect the epidemiological and demographic parameters.

Key words: Leshmaniasis, sandflyes, asymptomatics, Basic Reproduction Number
MSC 2000: AMS codes 92D30

1 Introduction

Leishmaniasis is a disease caused by at least two dozen of distinct protozoan of the Leishmania species, and it is transmitted by the bite of infected females of phlebotomine sandflies.

Depending on parasites, and the host immune response, Leshmaniasis can be subclinical (inapparent), localised (skin lesions), or can present a disseminated infection (cutaneous, mucosal, or visceral). The disease is associated with malnutrition, poor housing, and population displacement, among other factors. In general, it is estimated around 900 000 to 1.3 millon of new cases each year, and approximately 20 000 to 30 000 deaths occur worldwide each year [12].

Around 70 animal species, including humans, have been found as natural reservoir hosts of Leishmania parasites, and the clinical characteristics of the disease vary by regions. In general, the habit of keeping dogs and other domestic animals inside the house is thought to promote human infection.

Infection by leshmaniasis does not invariably cause illness in the host. The variable course of disease after infection is thought to due to differences in the immune response of individuals, and it can remain asymptomatic from one month to several years. In fact, many infected dogs remain asymptomatic and never develop clinical symptoms. In endemic regions, it is estimated that only one of five infected dogs develops clinical symptoms [2].

Epidemiological control of Leshmaniasis recommended by WHO includes vector control, human treatment, and control or sacrifice of infected dogs [8, 12]. Mathematical models are important tools to understand several aspects of the disease spread, as well as to asses the effectiveness of control measures, or to suggest new measures. Several non mathematical studies on Leishamniasis infection have been focus to understand the role of domestic dogs upon the disease spread. Since a percentage of dogs is lifelong asymptomatic, it is important to study the impact of the asymptomatics and infective dogs on disease prevalence.

2 Formulation of the model

We consider the infective populations of humans, dogs and sandflies. The first two populations are divided into the following epidemiological classes:

- Asymptomatic infectious
- Infectious with symptoms

with A_i, and I_i, $i = 1, 2$ denoting the asymptomatic infectives, and the infected with symptoms of each specie.

The population of infective sandflies is denoted by I_v.

The dynamics of the disease is modelled by the following system of differential equations

$$\begin{aligned} \frac{dA_h}{dt} &= r_h b\beta_h \frac{I_v}{\bar{N}}(\bar{N}_h - A_h - I_h) - (\gamma_h + \eta_h + \delta_h + \mu_h)A_h \\ \frac{dI_h}{dt} &= (1 - r_h)b\beta_h \frac{I_v}{\bar{N}}(\bar{N}_h - A_h - I_h) + \delta_h A_h - (\lambda_h + \tau_h + \mu_h)I_h \end{aligned}$$

$$\begin{aligned}
\frac{dA_d}{dt} &= r_d b\beta_d \frac{I_v}{\bar{N}}(\bar{N}_d - A_d - I_d) - (\gamma_d + \eta_d + \delta_d + \mu_d)A_d \\
\frac{dI_d}{dt} &= (1-r_d) b\beta_d \frac{I_v}{\bar{N}}(\bar{N}_d - A_d - I_d) + \delta_d A_d - (\lambda_d + \tau_d + \mu_d)I_d \\
\frac{dI_v}{dt} &= \frac{b}{\bar{N}}\left(\alpha_{ah}A_{ah} + \alpha_h I_h + \alpha_{ad}A_{ad} + \alpha_d I_d\right)(\bar{N}_v - I_v) - \mu_v I_v
\end{aligned} \tag{1}$$

in the positive invariant region

$$\Omega = \{(A_h, I_h, A_d, I_d, I_v) \in R_+^5 | A_h + I_h \leq \bar{N}_h, A_d + I_d \leq \bar{N}_d, I_v \leq \bar{N}_v\}.$$

In the model, $\bar{N}_h$, $\bar{N}_d$, and $\bar{N}_v$ denote the population sizes of humans, dogs and vectors, respectively. All the populations are constant with mortality rates given by μ_h, μ_d, μ_v, respectively. Following [5] it is assumed that susceptible humans and dogs get infected by infectious sandflies at rates $b\beta_h\frac{\bar{I}_v}{\bar{N}}$ and $b\beta_d\frac{\bar{I}_v}{\bar{N}}$, respectively, where b is the biting rate of sandflies, β_i, $i = h, d$ are the probabilities that an infected bite gives rise to a new case in the respective populations, and $\bar{N} = \bar{N}_h + \bar{N}_v$. A fraction r_i, $i = h, d$ of each species becomes asymptomatic infectives, and $1 - r_i$ pass directly to the infectious class. Reciprocally, sandflies get infected from asymptomatic humans and dogs, and infective humans and dogs at rates $\alpha_{ai}b\frac{A_i}{\bar{N}}$, $\alpha_i b\frac{I_i}{\bar{N}}, i = h, d$, respectively, where α_i, and $\alpha_{ai}, i = h, d$ are the transmission probabilities from humans and dogs to mosquitoes. It is reasonable to assume that a symptomatic infective is more infectious that an asymptomatic one, therefore in this work we will assume that $\alpha_{ai} = f_i\alpha_i$, $0 \leq f_i \leq 1, i = h, d$.

The asymptomatic members of each species become infectious with symptoms at a rate δ_i, or return to the susceptible class because they recover or are treated at rates γ_i, and η_i, respectively with $i = h, d$. Analogously, infectious, I_h, and I_d recover and return to the susceptible class, or are treated at rates λ_i, τ_i, respectively with $i = h, d$.

3 Analysis of the model

The analysis of system (1) will be given in terms of the *Basic Reproduction Number*, R_0, which represents the average number of secondary infections caused by an infected individual during the infection period in a whole susceptible population.

Using the next generation operator approach [4, 10], we compute the Basic Reproduction number R_0 in terms of the epidemiological and demographic parameters:

$$R_0 = \sqrt{R_{h_0}^2 \frac{\bar{N}_h}{\bar{N}} + R_{d_0}^2 \frac{\bar{N}_d}{\bar{N}}} \tag{2}$$

 ISBN: 978-84-608-6082-2

where

$$R_{0_i} = \sqrt{b^2\beta_i\left(\frac{(1-r_i)\alpha_i}{\lambda_i+\tau_i+\mu_i}+\frac{r_i}{\gamma_i+\eta_i+\delta_i+\mu_i}(\alpha_{ai}+\alpha_i\frac{\delta_i}{\lambda_i+\tau_i+\mu_i})\right)\frac{\bar{N}_v}{\bar{N}}} \tag{3}$$

$i = h, d$ represent the number of secondary infections derived from an infected individual in the human-vector cycle, and dog-vector cycle, respectively.

System (1) has the equilibrium $E_0 = (0, 0, 0, 0, 0)$ called the *disease-free state*. Using Theorem 2 of [10] we have the following result in terms of R_0.

Theorem 1. If $R_0 < 1$, $E_0 = (0, 0, 0, 0, 0)$ is the only equilibrium in Ω, and it is globally asymptotically stable. If $R_0 > 1$, E_0 becomes unstable.

An *endemic state* is a non trivial equilibrium of system (1). We have the following result.

Theorem 2. When $R_0 > 1$, system (1) has a unique endemic state $E_1 = (A_h^*, I_h^*, A_d^*, I_d^*, I_v^*)$.

Global stability of E_1 in Ω was proved in the case $\lambda_i + \tau_i = \gamma_i + \eta_i$, $i = h, d$ using the Lyapunov function [11].

$$\begin{aligned} W = {} & A\left(S_h - S_h^* - S_h^*\ln\frac{S_h}{S_h^*}\right) + B\left(A_h - A_h^* - A_h^*\ln\frac{A_h}{A_h^*}\right) + C\left(I_h - I_h^* - I_h^*\ln\frac{I_h}{I_h^*}\right) \\ & + D\left(S_d - S_d^* - S_d^*\ln\frac{S_d}{S_d^*}\right) + E\left(A_d - A_d^* - A_d^*\ln\frac{A_d}{A_d^*}\right) + F\left(I_d - I_d^* - I_d^*\ln\frac{I_d}{I_d^*}\right) \\ & + G\left(I_v - I_v^* - I_v^*\ln\frac{I_v}{I_v^*}\right), \end{aligned}$$

where $S_i = N_i - A_i - I_i$, $i = h, d$, and A, B, C, D, E, F, G are constants.

4 Sensitivity Analysis

The model formulated in this work has numerous parameters whose estimations and references are given in Table 1. However, uncertainties are expected to arise in those estimates. It is illustrative to investigate the sensitivity of the Basic Reproduction Number, R_0, to changes of the parameters related to asymptomatic disease and infective dogs in order to see the effect of these two factors on the disease prevalence. For this end we calculate the differential of the expression of R_0 given in (2),

 ISBN: 978-84-608-6082-2

parameter	meaning	value
b	sandfly biting rate	0.79 day^{-1}
β_h	sandfly-human transmission	0.3
β_d	sandfly-dog transmission	0.25
α_h	human-sandfly transmission	0.14
α_d	dog-sandfly transmission	0.25
δ_h	transition rate from asymptomatic to infectious humans	0.001-0.03 day^{-1} 0.0155 day^{-1}
δ_d	transition rate from asymptomatic to infectious dogs	0.0003-0.03 day^{-1} 0.0151 day^{-1}
r_h	proportion of asymptomatic infections in humans	0.17
r_d	proportion of asymptomatic infections in dogs	0.8
γ_h	infected humans recovery rate	0.0056 day^{-1}
γ_d	infected dogs recovery rate	0.0056 day^{-1}
λ_h	asymptomatic humans recovery rate	0.0098 day^{-1}
λ_d	asymptomatic dogs recovery rate	0.0098 day^{-1}
μ_v	average sandflies mortality rate	0.03 day^{-1}
μ_d	average dogs mortality rate	0.0003 day^{-1}
μ_h	average human mortality rate	0.00004 day^{-1}

Table 1: Demographic and epidemiological parameters. Values are taken from [1, 2, 3, 6, 7, 8, 9]

$$\Delta R_0 \approx dR_0 = \frac{\partial R_0}{\partial p}\Delta p, \tag{4}$$

to approximate the variations of R_0 with respect the parameter p, where p denotes the asymptomatic fractions $(r_i, i = h, d)$ and the increment of dog population (N_d). We measure the variation of R_0 when p increases taking $\Delta p > 0$.

When the proportion r_i of species i, $i = h, d$, is increasing by a factor θ,

$$\frac{\partial R_0}{\partial r_i} = \frac{b^2\beta_i}{2R_0}\left(\frac{-\alpha_i}{\lambda_i+\tau_i+\mu_i} + \frac{1}{\gamma_i+\eta_i+\delta_i+\mu_i}(\alpha_{ai} + \alpha_i\frac{\delta_i}{\lambda_i+\tau_i+\mu_i})\right)\frac{\bar{N}_v}{\bar{N}}$$

This expression is bigger than zero if

$$\frac{\alpha_i}{\lambda_i+\tau_i+\mu_i}\left(\frac{\gamma_i+\eta_i+\mu_i}{\gamma_i+\eta_i+\delta_i+\mu_i}\right) < \frac{\alpha_{ai}}{\gamma_i+\eta_i+\delta_i+\mu_i} \tag{5}$$

The inequality implies that R_0 increases if the number of infections produced by an asymptomatic individual during its asymptomatic period is bigger than the number of infections

 ISBN: 978-84-608-6082-2

produced by a symptomatic one during its infectious period. If the opposite inequality holds, R_0 decreases. Assuming $\alpha_i = \alpha_{ai}$, inequality (5) depends only of the infective and asymptomatic period considering that both types of infected may recover or be cured. In this case,

$$\frac{1}{\gamma_i + \eta_i + \mu_i} > \frac{1}{\lambda_i + \tau_i + \mu_i} \tag{6}$$

Figure 1 shows the evolution of R_0 when the proportion of asymptomatic humans increases by a factor θ. In all the simulations R_0 increases linearly, but their initial values and corresponding increments change according the percentage of individuals that receive treatment and are cured. In the simulations represented in figure 1a) there is not treatment at all. Only sixty percent of infective humans and thirty percent of infected dogs are treated in the simulation corresponding to 1b). In figure 1c), 90%, and 70% of the infectives are cured, 40%, and 25% of the infected and asymptomatic dogs are treated. Finally, in graph 1d), 0.90 and 0.70 of the infected humans and dogs are treated, while 0.80 and 0.50 of the asymptomatic humans and dogs receives treatment.

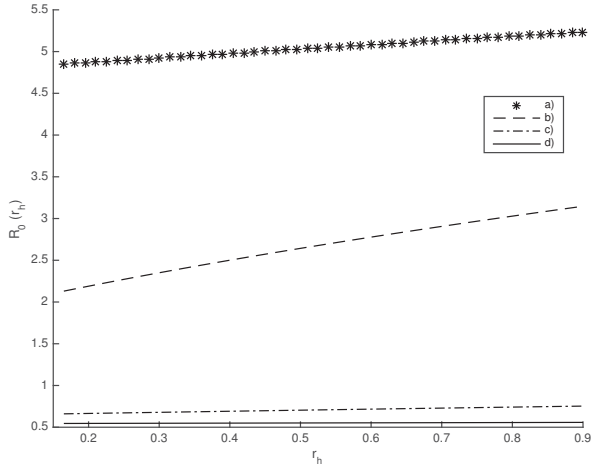


Figure 1

With no treatment, R_0 practically remains constant when the asymptomatics of each species increase. On the other hand, when treatment is applied only to infected ones, R_0 has a significant relative increase that seems contradictory. This behaviour can be explained observing that in this case R_0 has an important decrease at the beginning due to the treatment, but as the proportion of asymptomatics with respect to infectives, $r/(1-r)$, grows and infectives are treated, the number of infectious contacts with asymptomatics grow

rapidly. This increase coupled with the fact that the asymptomatic period can be very long, especially in dogs, gives rise to an increment of the number of secondary infections.

The behaviour described above is also observed when the proportion of asymptomatic dogs is increased.

If dog population N_d increments by a factor θ

$$\frac{\partial R_0}{\partial N_d} = \frac{1}{R_0(\bar{N}_h + \bar{N}_d)^3}(-2R_{0_h}N_h + R_{0_d}(N_h - N_d)),$$

then $\dfrac{\partial R_0}{\partial N_d} > 0$ if the ratios of dogs to humans, and R_{0_h} to R_{0_d} satisfy

$$1 - \frac{2R_{0_h}}{R_{0_d}} > \frac{N_d}{N_h}. \tag{7}$$

From this inequality we see that a necessary condition for R_0 to increase is that $R_{0_h} < \dfrac{R_{0_d}}{\sqrt{2}}$.

Figure 2 illustrates the behaviour of R_0 when dogs population increases, assuming the same treatment scheme as in figure 1. R_0 first increases, and then decreases approaching to a limit value due to saturation effect.

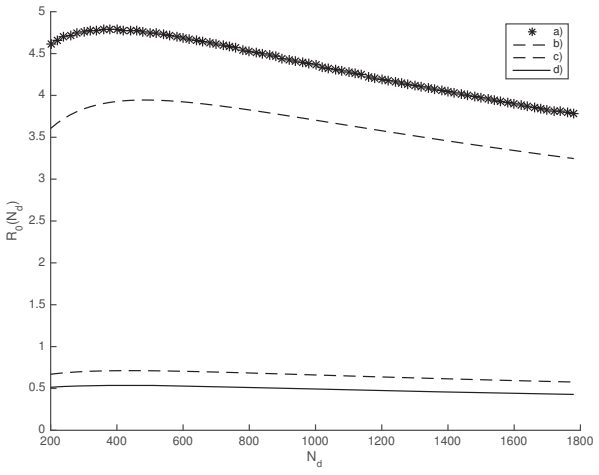


Figure 2

 ISBN: 978-84-608-6082-2

5 Conclusions

In this work we formulated a mathematical model in order to asses the impact of asymptomatic infectives and dogs in the transmission of Leshmaniasis disease. For this end we calculate the Basic Reproduction Number of the disease, R_0, and we made sensitive analysis of this parameter with respect the proportion of asymptomatics and dog population size. Our conclusions can be summarized as:

- R_0 increases if the number of infections produced by an asymptomatic individual during its asymptomatic period is bigger than the number of infections produced by a symptomatic one during its infectious period.
- Identification and treatment of asymptomatics of both species significantly reduces the prevalence of the disease.
- It is required a systematic application of serological tests in the regions where the disease is endemic to identify asymptomatic carriers and treat them.
- Depending on the ratios $\frac{R_{0_h}}{R_{0_d}}$ and $\frac{N_h}{N_d}$, R_0 increases or decreases when dog population grows.

Acknowledgements

This work has been partially supported by Projects IN-112713, IN113716 PAPIIT-UNAM.

References

[1] E.O. Agyingi, D.S. Ross, K. Bathena, *A model of the transmission dynamics of leishmaniasis*, J. Biol. Syst. **19** (2011) 237–250.

[2] Companion Animal Parasite Consul (CAPC), *Canine Leishmaniasis*, http:www.capcvet.org/capc-recommendations/canine-leishmaniasis.

[3] C.R. Davies, E.A. Llanos-Cuentas, S.D. M. Pyke, C. Dye, *Cutaneous leishmaniasis in the Peruvian Andes: an epidemiological study of infection and immunity*, Epidemiol. Infect. **114** (1995) 297–318.

[4] O. Diekmann, J.A.P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, Wiley, New York, NY, 2000.

[5] L. Esteva, C. Vargas, *Analysis of a dengue disease transmission model*, Math. Biosci. **150** (1998) 131-151.

[6] N. Hussaini, J.M.S Lubuma, K. Barley, A.B. Gumel, *Mathematical analysis of a model for AVL-HIV co-endemicity*, Math. Biosc. **271** (2016) 80–95.

[7] C.R. Palatnik-de-Souza, L.M. Bautista-de-Melo, G.P. Borja-Cabrera, M. Palatnik, C.L. Lavor, *Improving methods for epidemiological control of canine visceral leshmaniasis based on a mathematical model. Impact on the incidence of the canine and human disease*, Anais da Academia Brasileira de Ciências **76** (2004) 583–593.

[8] Panamerican Health Organization: *Control of Communicable Diseases Manual, 17th Edition*, James Chin Ed., American Public Health Association, 2000.

[9] S.E. Shaw, M.J. Day, *Artrophod-borne Infectious Diseases of the Dog and Cat*, CRC Press, Boca Ratón, Florida, USA, 2005.

[10] P. van den Driessche, P.J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartment models of disease transmission*, Math. Biosci. **180** (2002) 29-48.

[11] C. Vargas-De-León, *Global analysis of a delayed vector-bias model for malaria transmission with incubation period in mosquitoes*, MBE **9** (2012) 165–174.

[12] WHO: *Leishmaniasis*, Fact sheet, http:www.who.int/mediacentre/factsheets/fs375/en/, updated 2016.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A faithful functor among algebras and graphs

Óscar J. Falcón[1], Raúl M. Falcón[2], Juan Núñez[1], Ana M. Pacheco[3] and María Trinidad Villar[1]

[1] *Department of Geometry and Topology, University of Seville (Spain)*

[2] *Department of Applied Mathematics I, University of Seville (Spain)*

[3] *Department of Quantitative Methods, Loyola University Andalusia (Spain)*

emails: oscfalgan@yahoo.es, rafalgan@us.es, jnvaldes@us.es, ampacheco@uloyola.es, villar@us.es

Abstract

The problem of identifying a functor between the categories of algebras and graphs is currently open. Based on a known algorithm that identifies isomorphisms of Latin squares with isomorphism of vertex-colored graphs, we describe here a pair of graphs that enable us to find a faithful functor between finite-dimensional algebras over finite fields and these graphs.

Key words: graph theory, algebra, finite field, isomorphism, isotopism, invariant.

1 Introduction

Graph Theory has revealed to be an interesting tool to deal with distinct aspects on the study of algebras [3, 4, 7, 9]. Nevertheless, the problem of identifying a functor that relates the category of algebras with that of graphs remains still open. Both categories are referred with respect to their corresponding isomorphisms among algebras and graphs. Based on a proposal of McKay et al. [10] for identifying isomorphisms of Latin squares with isomorphism of vertex-colored graphs, we describe here a pair of graphs that enable us to find a faithful functor between finite-dimensional algebras over finite fields and these graphs. We focus in particular on the distribution of partial-magma algebras into isomorphism classes by means of some isomorphism invariants related to the mentioned graphs.

2 Preliminaries

2.1 Isotopisms of algebras

In 1942, Albert [1] introduced the concept of isotopism of algebras as a generalization of that of isomorphism. Specifically, two n-dimensional algebras $(A, \cdot)$ and $(A', \circ)$ defined over the same field $\mathbb{K}$ are said to be *isotopic* if there exist three non-singular linear transformations f, g and h from A to A' such that

$$f(u) \circ g(v) = h(u \cdot v), \text{ for all } u, v \in A. \tag{1}$$

Hereafter, in order to simplify the notation and whenever no confusion arises, we do not write explicitly the products $\cdot$ and $\circ$. That is, we write the previous identity as $f(u)g(v) = h(uv)$, for all $u, v \in A$. The triple (f, g, h) is an *isotopism* between the algebras A and A'.

Let A be an n-dimensional algebra over a field $\mathbb{K}$ and let $\{e_1, \ldots, e_n\}$ be a basis of this algebra. The *structure constants* of A are the numbers $c_{ij}^k \in \mathbb{K}$ such that

$$e_i e_j = \sum_{k=1}^{n} c_{ij}^k e_k, \text{ for } 1 \leq i, j \leq n. \tag{2}$$

If the structure constants of an algebra are all of them zeros, then this algebra is called *abelian* .

Lemma 1. *The n-dimensional abelian algebra is not isotopic to any other n-dimensional algebra.* □

Let S be a vector subspace of an algebra A. The *left* and *right annihilators* of S in A are respectively defined as the sets

$$\mathrm{Ann}_{A^-}(S) = \{u \in A \mid uv = 0, \text{ for all } v \in S\}. \tag{3}$$

$$\mathrm{Ann}_{A^+}(S) = \{u \in A \mid vu = 0, \text{ for all } v \in S\}. \tag{4}$$

The intersection of both sets is called the *annihilator* of S in A. It is defined as

$$\mathrm{Ann}_A(S) = \{u \in A \mid uv = vu = 0, \text{ for all } v \in S\}. \tag{5}$$

Lemma 2. *Let (f, g, h) be an isotopism between two n-dimensional algebras A and A'. Let S be a vector subspace of A. Then,*

a) $f(\mathrm{Ann}_{A^-}(S)) = \mathrm{Ann}_{A'^-}(g(S))$.

b) $g(\mathrm{Ann}_{A^+}(S)) = \mathrm{Ann}_{A'^+}(f(S))$.

c) $f(\mathrm{Ann}_{A^-}(S)) \cap g(\mathrm{Ann}_{A^+}(S)) = \mathrm{Ann}_{A'}(f(S) \cap g(S))$. □

Proposition 1. *Let (f, g, h) be an isotopism between two n-dimensional algebras A and A'. Then,*

a) $f(\mathrm{Ann}_{A^-}(A)) = \mathrm{Ann}_{A'^-}(A')$.

b) $g(\mathrm{Ann}_{A^+}(A)) = \mathrm{Ann}_{A'^+}(A')$.

c) $f(\mathrm{Ann}_{A^-}(A)) \cap g(\mathrm{Ann}_{A^+}(A)) = \mathrm{Ann}_{A'}(A')$.

Proof. The result follows straightforward from Lemma 2 and the regularity of f and g. □

Hereafter, given a vector subspace S of an algebra A, we define the vector subspace $SA = \{uv \mid u \in S \text{ and } v \in A\}$. The *derived algebra* of the algebra A is then defined as the subalgebra

$$A^2 = AA = \{uv \mid u, v \in A\} \subseteq A. \tag{6}$$

Lemma 3. *Let (f, g, h) be an isotopism between both algebras A and A'. Then, $h(A^2) = A'^2$ and $\dim(A^2) = \dim(A'^2)$.* □

2.2 Partial-magma algebras

A *partial magma* is a finite set endowed with a partial binary operation. Hereafter, we suppose this set to be $[n] = \{1, \ldots, n\}$ and we denote the operation as $\cdot$. In this case, n is the *order* of the partial magma. Two partial magmas $([n], \cdot)$ and $([n], \circ)$ are said to be *isotopic* if there exist three permutations α, β and γ in the symmetric group S_n such that

$$\alpha(i) \circ \beta(j) = \gamma(i \cdot j), \text{ for all } i, j \leq n \text{ such that } i \cdot j \text{ exists.} \tag{7}$$

If $\alpha = \beta = \gamma$, then the partial magmas are said to be *isomorphic*. The triple (α, β, γ) constitutes an *isotopism* of magmas (an *isomorphism* if $\alpha = \beta = \gamma$).

A *partial quasigroup* is a partial magma $([n], \cdot)$ such that if the equations $ix = j$ and $yi = j$, with $i, j \in [n]$, have solutions for x and y in $[n]$, then these solutions are unique. Every partial quasigroup of order n is the multiplication table of a *partial Latin square* of order n, that is, an $n \times n$ array in which each cell is either empty or contains one element chosen from the set $[n]$, such that each symbol occurs at most once in each row and in each column. Every isotopism of a partial quasigroup is uniquely related to a permutation of the rows, columns and symbols of the corresponding partial Latin square. The distribution

 ISBN: 978-84-608-6082-2

of partial Latin squares into isotopism classes is known for order up to 6 [5, 6]. Finally, if two Latin squares are isotopic after a reordering of the components of all their entries, then they are said to be *paratopic*.

In 1944, Bruck [2] introduced the concept of *quasigroup algebra* as an n-dimensional algebra over a base field $\mathbb{K}$ such that there exists a basis $\{e_1, \ldots, e_n\}$ and a quasigroup $([n], \cdot)$ satisfying that $e_i e_j = c_{ij} e_{i \cdot j}$ for each pair of elements $i, j \leq n$ and some non-zero structure constant $c_{ij} \in \mathbb{K} \setminus \{0\}$. The algebra is then said to be *based on* the quasigroup $([n], \cdot)$. If all its structure constants are equal to 1, then this is called a *quasigroup ring*. *Partial-magma algebras* constitute a natural generalization of the concept of quasigroup algebra, once the condition of being based on a quasigroup is replaced by that of being based on a partial magma.

2.3 Graph theory

A *graph* is a pair $G = (V, E)$ formed by a set V of points or *vertices* and a set E of lines or *edges* formed by subsets of two vertices of V. The *degree* of a vertex $v \in V$ is the number $d(v)$ of edges containing this vertex. A graph is said to be *vertex-colored* if there exists a partition into color sets of its set of vertices. The color of a vertex v is denoted as $\text{color}(v)$. An *isomorphism* between two vertex-colored graphs $G = (V, E)$ and $G' = (V', E')$ is any bijective map f between the set of vertices V and V' that preserves collinearity and such that $\text{color}(f(v)) = \text{color}(v)$, for all $v \in V$.

Let $L = (l_{ij})$ be a Latin square of order n. McKay et al. [10] defined the vertex-colored graph $G_2(L)$ with $n^2 + 3n$ vertices

$$\{r_i \mid i \leq n\} \cup \{c_i \mid i \leq n\} \cup \{s_i \mid i \leq n\} \cup \{t_{ij} \mid i, j \leq n\},$$

where each of the four subsets (related to the rows (r_i), columns (c_i), symbols (s_i) and cells (t_{ij}) of the Latin square L) has a different color, and $3n^2$ edges

$$\{r_i e_{ij}, c_j e_{ij}, s_{ij} t_{ij} \mid i, j \leq n\}\}.$$

They also defined the vertex-colored graph $G_1(L)$ from the graph $G_2(L)$ by adding 3 additional vertices $\{R, C, S\}$ and $3n$ additional edges $\{Rr_i, Cc_i, Ss_i \mid i \leq n\}$. Here, there are three colors: one for $\{R, C, S\}$, one for $\{r_i, c_i, s_i \mid i \leq n\}$ and one for the rest of vertices. Finally, they defined the vertex-colored graph $G_3(L)$ from the graph $G_2(L)$ by adding $3n$ additional edges $\{r_i c_i, c_i s_i, r_i s_i \mid i \leq n\}$. Here, the color of the vertices coincides with those of $G_1(L)$. These authors proved (Theorem 6 in [10]) that two Latin squares L_1 and L_2 of the same order are paratopic (respectively, isotopic or isomorphic) if and only if the graphs $G_1(L_1)$ and $G_1(L_2)$ (respectively, $G_2(L_1)$ and $G_2(L_2)$, and $G_3(L_1)$ and $G_3(L_2)$) are

isomorphic. Figure 1 shows an example of the three graphs related to the next Latin square of order 2.

$$L = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

We have used distinct styles ($\circ$, $\blacktriangle$, $\blacktriangleright$, $\blacktriangleleft$ and $\bullet$) in the vertices of the graphs to represent their colors.

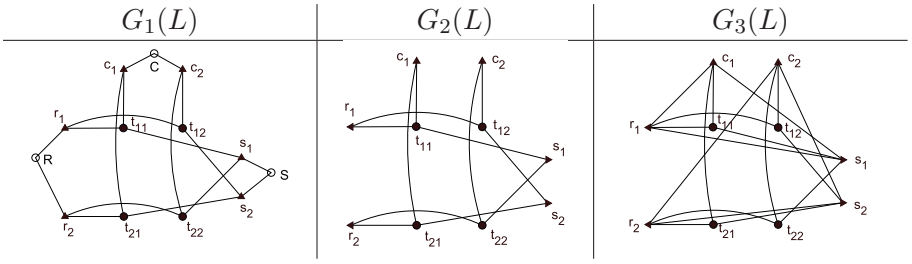


Figure 1: Graphs related to a Latin square of order 2.

3 The proposed graph

Based on the proposal of McKay et al. for Latin squares, we describe now a pair of graphs that are uniquely related to a finite-dimensional algebra over a finite field and which enable us to ensure that any two isotopic or isomorphic algebras map to two isomorphic graphs. To this end, let A be an n-dimensional algebra over a finite field $\mathbb{K}$. Firstly, we define the vertex-colored graph $G_1(A)$ with four maximal monochromatic subsets

$$\begin{cases} R_A = \{r_u \mid u \in A \setminus \mathrm{Ann}_{A^-}(A)\}, \\ C_A = \{c_u \mid u \in A \setminus \mathrm{Ann}_{A^+}(A)\}, \\ S_A = \{s_u \mid u \in A^2 \setminus \{0\}\}, \\ T_A = \{t_{u,v} \mid u, v \in A, uv \neq 0\}. \end{cases}$$

and edges

$$\{r_u t_{u,v}, c_v t_{u,v}, s_w t_{u,v} \mid u, v, w \in A, uv = w \neq 0\}.$$

From this graph we also define the vertex-colored graph $G_2(A)$ by adding the edges

$$\{r_u c_u, \mid u \in A \setminus \mathrm{Ann}_A(A)\} \cup \{c_u s_u \mid u \in A^2 \setminus \mathrm{Ann}_{A^+}(A)\} \cup \{r_u s_u \mid u \in A^2 \setminus \mathrm{Ann}_{A^-}(A)\}.$$

Figure 2 shows, for instance, the two graphs G_1 and G_2 that are related to any n-dimensional anticommutative algebra over the finite field $\mathbb{F}_2$, with basis $\{e_1, \ldots, e_n\}$, that is described by the non-zero product $e_1 e_2 = e_1$.

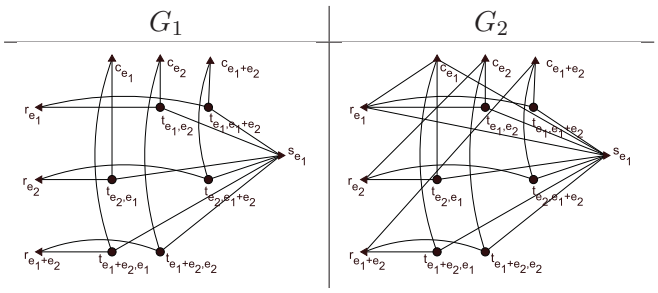


Figure 2: Graphs related to the anticommutative algebra $e_1e_2 = e_1$ over $\mathbb{F}_2$.

Example 1. *In order to illustrate in a better way the proposed graphs, we describe how to construct step by step the graph G_2 related to the 3-dimensional anti-commutative algebra A over the finite field $\mathbb{F}_2$, with basis $\{e_1, e_2, e_3\}$, that is linearly defined from the non-zero products $e_1e_3 = e_2$ and $e_2e_3 = e_1$. In order to make easier this construction, we place the vertices of the graph in rows and columns as if they were the elements of a matrix. Each one of these vertices can, therefore, be described by its position (i, j) inside this matrix.*

Step 1. The underlying set of vectors of our algebra is $\{e_1, e_2, e_3, e_1+e_2, e_1+e_3, e_2+e_3, e_1+e_2+e_3\}$. Since $\mathrm{Ann}_A(A) = \emptyset$, these seven vectors can appear as left or right factors of a non-zero product in A. We start, therefore, the construction of the graph $G_2(A)$ by drawing seven vertices labeled as r_u in a column at the left of the graph and other seven vertices labeled as c_u in a row on the top of the graph. Each u denotes here a vector of the algebra (see Figure 3 (left)).

Step 2. In the body of the graph (the empty zone among the two sets of vertices that have been drawn until now) we draw now those vertices corresponding to non-null brackets. Since we have at most seven times seven brackets, we add at most 49 new vertices labeled as $t_{u,v}$. These vertices are distributed in matrix form according to the left and right factors that determine the corresponding product (see Figure 3 (center)).

Step 3. Now, since $A^2 \setminus \{0\} = \{e_1, e_2, e_1 + e_2\}$, we draw three new vertices labeled as s_u in a column at the right of the graph (see Figure 3 (right)).

Step 4. We deal now with the construction of the edges. Firstly, we join each vertex $(i, 1)$ at the left of the graph with the vertices (i, j) for $2 \leq i, j \leq 7$. These correspond to the edges $r_ut_{u,v}$ of the description (see Figure 4 (left)).

Step 5. After that, we join each vertex $(1, j)$ on the top of the graph with the vertices (i, j) for $2 \leq i, j \leq 7$. These correspond to the edges $c_ut_{u,v}$ of the description (see Figure 4 (left in the center)).

Figure 3: Steps 1–3 in the construction of the graph $G_2(A)$.

Step 6. Next, we join each vertex $(i, 1)$ *with the vertex* $(1, i)$ *for* $2 \leq i \leq 7$*. These correspond to the edges* $r_u c_u$ *of the description. (see Figure 4 (center)).*

Step 7. Now, we join each of the vertices $t_{u,v}$ *with the corresponding vertex constructed in Step 3. These correspond to the edges* $s_w t_{u,v}$ *of the description (see Figure 4 (right in the center)).*

Step 8. Finally, whenever is possible, we join the vertices $(1, j)$ *and* $(i, 1)$ *with the corresponding vertices constructed in Step 3, for* $2 \leq i, j \leq 7$*. These correspond, respectively, to the edges* $r_u s_u$ *and* $c_u s_u$ *of the description (see Figure 4 (right)).* $\lhd$

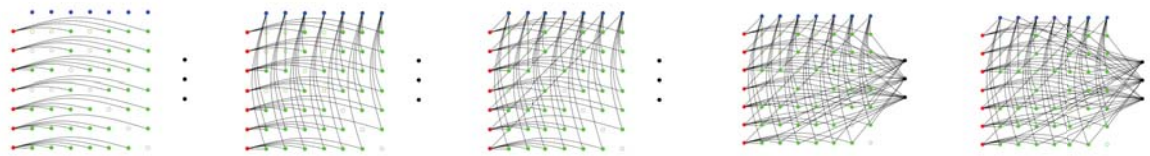


Figure 4: Steps 4–8 in the construction of the graph $G_2(A)$.

Lemma 4. *Let* A *be an* n*-dimensional algebra over a finite field* $\mathbb{K}$*. Then,*

a) If the algebra A *is abelian, then both graphs* $G_1(A)$ *and* $G_2(A)$ *are empty.*

b) The graph $G_1(A)$ *does not contain triangles.*

c) In both graphs $G_1(A)$ *and* $G_2(A)$*,*

- *The number of vertices is*

$$|A \setminus \mathrm{Ann}_{A^-}(A)| + |A \setminus \mathrm{Ann}_{A^+}(A)| + |A^2| + |\{(u, v) \in A \times A \mid uv \neq 0\}| - 1.$$

- $d(t_{u,v}) = 3$, *for all* $u, v \in A$ *such that* $uv \neq 0$.

d) In the graph $G_1(A)$,

- $d(r_u) = |A \setminus \mathrm{Ann}_{A^+}(\{u\})|$, *for all* $u \in A \setminus \mathrm{Ann}_{A^-}(A)$.
- $d(c_u) = |A \setminus \mathrm{Ann}_{A^-}(\{u\})|$, *for all* $u \in A \setminus \mathrm{Ann}_{A^+}(A)$.
- $d(s_u) = \sum_{v \in A} |\mathrm{ad}_v^{-1}(u)|$, *for all* $u \in A^2 \setminus \{0\}$. *Here,* ad *denotes the adjoint action.*

e) In the graph $G_2(A)$,

- $d(r_u) = |A \setminus \mathrm{Ann}_{A^+}(\{u\})| + \mathbf{1}_{A \setminus \mathrm{Ann}_{A^-}(A)}(u) + \mathbf{1}_{A^2}(u)$, *for all* $u \in A \setminus \mathrm{Ann}_{A^-}(\{u\})$.
- $d(c_u) = |A \setminus \mathrm{Ann}_{A^-}(\{u\})| + \mathbf{1}_{A \setminus \mathrm{Ann}_{A^+}(A)}(u) + \mathbf{1}_{A^2}(u)$, *for all* $u \in A \setminus \mathrm{Ann}_{A^+}(\{u\})$.
- $d(s_u) = \mathbf{1}_{A \setminus \mathrm{Ann}_{A^-}(A)}(u) + \mathbf{1}_{A \setminus \mathrm{Ann}_{A^+}(A)}(u) + \sum_{v \in A} |\mathrm{ad}_v^{-1}(u)|$, *for all* $u \in A^2 \setminus \{0\}$.

Here, $\mathbf{1}$ *denotes the characteristic function.*

Proposition 2. *Let* A *be an* n*-dimensional algebra over a finite field* $\mathbb{K}$. *Then,*

a) The number of edges of its related graph $G_1(A)$ *is*

$$\sum_{u \notin \mathrm{Ann}_{A^-}(A)} (|A \setminus \mathrm{Ann}_{A^+}(\{u\})| + \sum_{v \in A^2 \setminus \{0\}} |\mathrm{ad}_u^{-1}(v)|) + \sum_{u \notin \mathrm{Ann}_{A^+}(A)} |A \setminus \mathrm{Ann}_{A^-}(\{u\})|.$$

b) The number of edges of its related graph $G_2(A)$ *coincides with those of* $G_1(A)$ *plus*

$$|A \setminus \mathrm{Ann}_A(A)| + |A^2 \setminus \mathrm{Ann}_{A^-}(A)| + |A^2 \setminus \mathrm{Ann}_{A^+}(A)|.$$

Proof. The result follows straightforward from the First Theorem of Graph Theory [8] and assertions (c–e) in Lemma 4. □

Theorem 1. *Let* A *and* A' *be two* n*-dimensional algebras over a finite field* $\mathbb{K}$. *Then,*

a) If both algebras are isotopic, then their corresponding graphs $G_1(A)$ *and* $G_1(A')$ *are isomorphic. Reciprocally, if the graphs* $G_1(A)$ *and* $G_1(A')$ *are isomorphic, then there exist three bijective maps* f, g *and* h *between* A *and* A' *such that* $f(u)g(v) = h(uv)$.

b) If both algebras are isomorphic, then their corresponding graphs $G_2(A)$ *and* $G_2(A')$ *are also isomorphic. Reciprocally, if the graphs* $G_2(A)$ *and* $G_2(A')$ *are isomorphic, then there exists a multiplicative bijective map between the algebras* A *and* A', *that is, a bijective map* $f : A \to A'$ *so that* $f(u)f(v) = f(uv)$, *for all* $u, v \in A$.

Proof. Let (f, g, h) be an isotopism between the algebras A and A'. We define the map α between $G_1(A)$ and $G_1(A')$ such that

$$\begin{cases} \alpha(r_u) = r_{f(u)}, \text{ for all } u \in A \setminus \mathrm{Ann}_{A^-}(A), \\ \alpha(c_u) = c_{g(u)}, \text{ for all } u \in A \setminus \mathrm{Ann}_{A^+}(A), \\ \alpha(s_u) = s_{h(u)}, \text{ for all } u \in A^2 \setminus \{0\}, \\ \alpha(t_{u,v}) = t_{f(u),g(v)}, \text{ for all } u, v \in A \text{ such that } uv \neq 0. \end{cases}$$

The description of both graphs $G_1(A)$ and $G_1(A')$ together with Proposition 1, Lemma 3 and the regularity of f, g and h involve α to be an isomorphism between these two vertex-colored graphs, that is, α is a well-defined bijection between the vertices of $G_1(A)$ and $G_1(A')$ that preserves collinearity and the color of the vertices. The same map α constitutes an isomorphism between the graphs $G_2(A)$ and $G_2(A')$ in case of being $f = g = h$, that is, if the algebras A and A' are isomorphic.

Reciprocally, let α be an isomorphism between the graphs $G_1(A)$ and $G_1(A')$. Collinearity involves this isomorphism to be uniquely determined by its restriction to $R_A \cup C_A \cup S_A$. Specifically, the image of each vertex $t_{u,v} \in T_A$ by means of α is uniquely determined by the corresponding images of r_u, c_v and s_{uv}. Let β and β' be the respective bases of the algebras A and A' and let $\pi : A \to A'$ be the natural map that preserves the components of each vector with respect to the mentioned bases. That is, $\pi((u_1, \ldots, u_n)_\beta) = (u_1, \ldots, u_n)_{\beta'}$, for all $u_1, \ldots, u_n \in \mathbb{K}$. Let us define three maps f, g and h from A to A' such that

$$f(u) = \begin{cases} \pi(u), \text{ for all } u \in \mathrm{Ann}_{A^-}(A), \\ v, \text{ otherwise, where } v \in A \text{ is such that } \alpha(r_u) = r_v. \end{cases}$$

$$g(u) = \begin{cases} \pi(u), \text{ for all } u \in \mathrm{Ann}_{A^+}(A), \\ v, \text{ otherwise, where } v \in A \text{ is such that } \alpha(c_u) = c_v. \end{cases}$$

$$h(u) = \begin{cases} \pi(u), \text{ for all } u \in (A \setminus A^2) \cup \{0\}, \\ v, \text{ otherwise, where } v \in A \text{ is such that } \alpha(s_u) = s_v. \end{cases}$$

From Proposition 1 and Lemma 3, these three maps are bijective. Let $u, v \in A$. If $u \in \mathrm{Ann}_{A^-}(A)$ or $v \in \mathrm{Ann}_{A^+}(A)$, then there does not exist the vertex $t_{u,v}$ in the graph $G_1(A)$. Since α preserves collinearity, there does not exist the vertex $t_{f(u),g(v)}$ in the graph $G_1(A')$, which means that $f(u) \in \mathrm{Ann}_{A'^-}(A')$ or $g(v) \in \mathrm{Ann}_{A'^+}(A')$. In any case, we have that $f(u)g(v) = 0 = h(uv)$. Finally, if $u \notin \mathrm{Ann}_{A^-}(A)$ and $v \notin \mathrm{Ann}_{A^+}(A)$, then the vertex $t_{u,v}$ connects the vertices r_u, c_v and s_{uv} in the graph $G_1(A)$. Now, the isomorphism α maps this vertex $t_{u,v}$ in $G_1(A)$ to a vertex $t_{u',v'}$ in $G_2(A)$ that is connected to the vertices $r_{u'}$, $c_{v'}$ and $s_{u'v'}$. Again, since α preserves collinearity, it is $f(u) = u'$, $g(v) = v'$ and, finally, $h(uv) = f(u)g(v)$.

In case of being α an isomorphism between the graphs $G_2(A)$ and $G_2(A')$ it is enough to consider $f = g = h$ in the previous description. This is well-defined because of the new edges that are included in the graphs $G_1(A)$ and $G_1(A')$ in order to define, respectively, the graphs $G_2(A)$ and $G_2(A')$. These edges involve the multiplicative character of the bijective map f, that is, $f(u)g(v) = h(uv)$, for all $u, v \in A$. □

Theorem 1 enables us to determine non-isotopic and non-isomorphic algebras from their corresponding non-isomorphic graphs. To this end, it is interesting to compute some isomorphism invariants of the corresponding graphs G_1 and G_2. In this regard, Table 1 shows, for instance, some graph invariants of the graph G_1 related to each one of the possible isomorphism classes of 3-dimensional Lie algebras over the finite fields $\mathbb{F}_2$ and $\mathbb{F}_3$. All of them constitute partial-magma algebras.

	$\mathbb{F}_2$			$\mathbb{F}_3$		
Lie partial-magma algebra	Vertices	Edges	Triangles	Vertices	Edges	Triangles
Abelian	1	0	0	1	0	0
$e_1e_2 = e_3$	37	72	0	482	1296	0
$e_1e_2 = e_2$	37	72	0	482	1296	0
$e_1e_2 = e_3, e_1e_3 = -e_2$	-	-	-	636	1728	0
$e_1e_2 = e_3, e_1e_3 = e_2$	53	108	0	636	1728	0
$e_1e_2 = e_2, e_1e_3 = e_3$	53	108	0	636	1728	0
$e_1e_2 = e_2, e_1e_3 = -e_3, e_2e_3 = -e_1$	63	126	0	-	-	-
$e_1e_2 = e_2, e_1e_3 = -e_3, e_2e_3 = 2e_1$	-	-	-	702	1872	0

Table 1: Graph invariants for the graph G_1 related to each isomorphism class of 3-dimensional Lie partial-magma algebras over the finite fields $\mathbb{F}_2$ and $\mathbb{F}_3$.

Thus, for instance, it is known that the n-dimensional anticommutative algebra over the finite field $\mathbb{F}_2$, with $n \geq 3$, described by the product $e_1e_2 = e_3$ is not isomorphic to the n-dimensional anticommutative algebra over $\mathbb{F}_2$ described by the product $e_1e_2 = e_1$. This follows straightforward from the fact that the corresponding graph G_2 related to the former coincides with that associated with the latter, which is shown in Figure 2 (right), up to the vertex s_{e_1}, which becomes s_{e_3}, and the two edges $r_{e_1}s_{e_1}$ and $c_{e_1}s_{e_1}$, which disappear. Both graphs are, therefore, non-isomorphic and hence, the algebras are neither isomorphic. It is straightforward verified that both algebras are, however, isotopic. Besides, even if this does not constitute a necessary condition, their corresponding graphs G_1 are isomorphic. That graph shown in Figure 2 (left) is indeed the graph G_1 corresponding to the anticommutative algebra described by the product $e_1e_2 = e_1$.

We finish this paper with an illustrative example that focuses on those graphs G_1 and G_2 related to the set of non-abelian partial-quasigroup rings over a finite field that are based on the known distribution of partial Latin squares of order 2 into isotopism classes.

 ISBN: 978-84-608-6082-2

Particularly, Table 2 shows the graph invariants related to the finite field $\mathbb{F}_2$. Partial Latin squares are written row after row in a single line, with empty cells represented by zeros. For each isotopism class we indicate the sequence with the number of vertices of each color, the number of edges and that of triangles of the corresponding graphs G_1 and G_2.

	G_1			G_2		
Partial Latin square	Vertices	Edges	Triangles	Vertices	Edges	Triangles
10 00	(2,2,1,4)	12	0	(2,2,1,4)	16	7
10 01	(3,3,1,6)	18	0	(3,3,1,6)	23	7
10 02	(3,3,3,7)	21	0	(3,3,3,7)	30	16
10 20	(3,2,3,6)	18	0	(3,2,3,6)	25	12
12 00	(2,3,3,6)	18	0	(2,3,3,6)	25	12
12 20	(3,3,3,8)	24	0	(3,3,3,8)	33	13
12 21	(3,3,3,8)	24	0	(3,3,3,8)	33	13

Table 2: Graph invariants for the graphs G_1 and G_2 related to 2-dimensional non-abelian partial-quasigroup rings over the finite field $\mathbb{F}_2$.

Theorem 2. *The set of* 2*-dimensional non-abelian partial-quasigroup rings is distributed into six isotopism classes.*

Proof. A computational case study enables us to ensure the result. In particular, if the characteristic of the base field is distinct of two, then the six isotopism classes under consideration are those related to the next partial Latin squares of order 2

$$\begin{array}{|c|c|}\hline 1 & \\ \hline & \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & \\ \hline & 1 \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & \\ \hline 2 & \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline & \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 2 & \\ \hline\end{array} \quad \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 2 & 1 \\ \hline\end{array}$$

Otherwise, if the characteristic of the base field is two, then the isotopism classes related to the last two partial Latin squares coincide. In this case, the next partial Latin square corresponds to the sixth isotopism class

$$\begin{array}{|c|c|}\hline 1 & \\ \hline & 2 \\ \hline\end{array}$$

If the characteristic of the base field is distinct of two, the partial-quasigroup ring related to this partial Latin square is isotopic to that related to the unique Latin square of the previous list. □

 ISBN: 978-84-608-6082-2

4 Conclusion and further studies

We have described in this paper a pair of graphs that enable us to define faithful functors between finite-dimensional algebras over finite fields and these graphs. The computation of isomorphism invariants of these graphs plays a remarkable role in the distribution of distinct families of algebras into isotopism and isomorphism classes. Some preliminary results have been exposed in this regard, particularly on the distribution of partial-quasigroup rings over finite fields. Based on the known classification of partial Latin squares into isotopism classes, further work is required to determine completely this distribution.

References

[1] Albert, A.A. *Non-Associative Algebras: I. Fundamental Concepts and Isotopy*, Ann. of Math., Second Series **43** (1942) 685–707.

[2] Bruck, R. H. *Some results in the theory of quasigroups*, Trans. Am. Math. Soc. **55** (1944), 19–52.

[3] Carriazo, A., Fernández, L. M. and Núñez, J. *Combinatorial structures associated with Lie algebras of finite dimension*, Linear Algebra Appl. **389** (2004) 43–61.

[4] Dani, S.G. and Mainkar, M.G. *Anosov automorphisms on compact nilmanifolds associated with graphs*, Trans. Amer. Math. Soc. **357**:6 (2005), 2235–2251.

[5] Falcón, R.M. *The set of autotopisms of partial Latin squares*, Discrete Math. **313** (2013), 1150–1161.

[6] Falcón, R.M. and Stones, R.J. *Classifying partial Latin rectangles*, Electron. Notes Discrete Math. **49** (2015), 765–771.

[7] Hamelink, R.C. *Graph theory and Lie algebra*. In: The Many Facets of Graph Theory, Proc. Conf., Western Mich. Univ., Kalamazoo, Mich., 1968, Springer, Berlin, 1969, pp. 149153.

[8] Harary, F. *Graph Theory*, Addison Wesley, Reading, Mass., 1969.

[9] Mainkar, M.G. *Graphs and two-step nilpotent Lie algebras*, Groups Geom. Dyn. **9**:1 (2015), 55–65.

[10] McKay, B. D., Meynert, A. and Myrvold, W. *Small Latin Squares, Quasigroups and Loops*, J. Combin. Des. **15** (2007), 98–119.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Trigonometrically fitted explicit symmetric six-step methods for undamped Duffing equation

Yonglei Fang[1]

[1] *School of Mathematics and Statistics, Zaozhuang University, Zaozhuang 277160, P.R. China*

emails: ylfangmath@163.com

Abstract

Multi-frequency trigonometrically fitted multistep methods for the numerical integration of the undamped Duffing equation are investigated. Three new trigonometrically fitted symmetric six-step methods containing different resonance spectrum are constructed. Numerical results are reported to show the efficiency of the new methods.

Key words: Explicit symmetric method, multi-frequency problems, undamped Duffing equation
MSC 2000: 65L05, 65L06

1 Introduction

The undamped Duffing cubic equation driven by a periodic force has the form

$$y''(x) + y(x) + y(x)^3 = B\cos(\omega x). \tag{1}$$

With the harmonic balance method, Mickens [1] showed that the solution of Equation (1) can be expanded in a series of periodic functions

$$y(x) = \sum_{i=0}^{\infty} A_{2i+1}\cos\left((2i+1)\omega x\right). \tag{2}$$

However, estimating of all the coefficients A_{2i+1} is a task of challenge. On the other hand, numerical solution of the undamped Duffing equation has attracted more and more interests.

 ISBN: 978-84-608-6082-2

Recently, with the Fourier spectrum, Wang [2] proposed trigonometrically fitted Numerov-type methods for the undamped Duffing equation for which the accuracy and stability have been greatly improved. More recently, Wang [3] constructed a new kind of trigonometrically fitted Obrechkoff one-step method, which combines the Fourier spectrum with the high-order derivatives. Unfortunately, Wang's methods are implicit, for which it needs to solve a nonlinear algebraic system at each step. Even if for the undamped Duffing equation, the computational cost is high. In this paper, we will derive a new family of explicit symmetric six-step methods which will be shown to be more efficient.

2 Construction of the new methods

The Duffing equation (1) can be written in a general form of second-order ODE

$$y''(x) = f(x, y), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0, \quad x \in [x_0, x_{\text{end}}]. \tag{3}$$

We consider the following explicit symmetric six-step method

$$y_{n+3} - y_{n+2} - y_{n-2} + y_{n-3} = h^2(b_0(f_{n+2} + f_{n-2}) + b_1(f_{n+1} + f_{n-1}) + b_2 f_n), \tag{4}$$

where $y_{n\pm i} = y(x \pm ih),\ f_{n\pm i} = y''(x \pm ih), i = -3, \cdots, 3$ (see Simos [4]). The frequency spectrum of the undamped Duffing equation (1) consists of odd multiples of ω and the resonant oscillation arises just at these frequencies. With the choice of

$$\{b_0, b_1, b_2\} = \{\frac{67}{48}, -\frac{1}{6}, \frac{61}{24}\},$$

we obtain the traditional explicit symmetric six-step method which can be found in Simos [4]. The local truncation error of this method is

$$LTER(h) = \frac{787}{12096} h^8 y^{(8)}(x) + \mathcal{O}(h^{10}).$$

Therefore this method is of order six and we denote it as Method I. In the sequel, in order to derive new methods adapted to oscillatory problems, we follow the line of Fang [5] and require the method (4) to integrate exactly functions in the set

$$\mathcal{F}_K = \{\sin(k\omega t), \cos(k\omega t),\ k = 1, \ldots, K\},\ \omega \in \mathbb{R}. \tag{5}$$

2.1 Method II

Requiring the method (4) to be exact for functions in the set $\mathcal{F}_1$ leads to

$$2(\cos(3u) - \cos(2u)) = -u^2(2(d_0 \cos(2u) + d_1 \cos(u)) + d_2), u = \omega h. \tag{6}$$

 ISBN: 978-84-608-6082-2

Fixing $b_1 = -\frac{1}{6},\ b_2 = \frac{61}{24}$ and solving the equation (6) gives

$$b_0 = (-48(\cos(3u) - \cos(2u)) + 8u^2\cos(u) - 61u^2)/(48u^2\cos(2u)).$$

The local truncation error of this new method is given by

$$LTER(h) = \frac{787}{12096}h^8(\omega^6 y^{(2)}(x) + y^{(8)}(x)) + \mathcal{O}(h^{10}).$$

Thus, this method is of order six and we denote it as Method II.

2.2 Method III

Requiring the method (4) to be exact for functions in the set $\mathcal{F}_2$ leads to

$$\begin{cases} 2(\cos(3u) - \cos(2u)) = -u^2(2(d_0\cos(2u) + d_1\cos(u)) + d_2), \\ 2(\cos(9u) - \cos(6u)) = -9u^2(2(d_0\cos(6u) + d_1\cos(3u)) + d_2). \end{cases} \tag{7}$$

With the choice of $b_2 = \frac{61}{24}$, we solve equations (7) and obtain

$$\begin{aligned} b_0 &= \csc(u)^2(-3(61u^2\cos(3u) + 48\cos(3u)^2 - 24\cos(5u)) \\ &\quad + \cos(u)(72 + 183u^2 - 16\cos(6u) + 16\cos(9u)))/M, \\ b_1 &= \csc(u)^2(-\cos(2u)(183u^2 + 128\cos(6u) + 16\cos(9u)) \\ &\quad + 3\cos(6u)(61u^2 + 48\cos(3u)))/M, \end{aligned}$$

where $M = 288u^2(3\cos(u) + 2\cos(3u) + \cos(5u)$. The local truncation error of this new method is given by

$$LTER(h) = -\frac{787}{12096}h^8(90\omega^6 y^{(2)}(x) + 91\omega^4 y^{(4)}(x) - y^{(8)}(x)) + \mathcal{O}(h^{10}).$$

So this method is of order six, and we denote it as Method III.

2.3 Method IV

Requiring the method (4) to be exact for functions in the set $\mathcal{F}_3$ leads to

$$\begin{cases} 2(\cos(3u) - \cos(2u)) = -u^2(2(d_0\cos(2u) + d_1\cos(u)) + d_2), \\ 2(\cos(9u) - \cos(6u)) = -9u^2(2(d_0\cos(6u) + d_1\cos(3u)) + d_2), \\ 2(\cos(15u) - \cos(10u)) = -25u^2(2(d_0\cos(10u) + d_1\cos(5u)) + d_2). \end{cases} \tag{8}$$

 ISBN: 978-84-608-6082-2

Solving the equations (8) gives

$$\begin{aligned}
b_0 &= (\csc(u)^6\sec(u)^3(-225\cos(3u)^2+25\cos(5u)(\cos(6u)-\cos(9u))\\
&\quad +9\cos(3u)(25\cos(5u)-\cos(10u)+\cos(15u))+\cos(u)(-25\cos(6u)\\
&\quad +25\cos(9u)+9\cos(10u)-9\cos(15u)+1800\cos(2u)^2\sin(u)^2)))/N,\\
b_1 &= ((225\cos(3u)(\cos(6u)-\cos(10u))-16\cos(6u)\cos(10u)\\
&\quad +\cos(2u)(-200\cos(6u)-25\cos(9u)+216\cos(10u)+9\cos(15u))\\
&\quad +25\cos(9u)\cos(10u)-9\cos(6u)\cos(15u))\csc(u)^6\sec(u)^3)/N,\\
b_2 &= ((2432+4504\cos(u)+3559\cos(2u)+2434\cos(3u)\\
&\quad +1579\cos(4u)+1064\cos(5u)+829\cos(6u)+734\cos(7u)\\
&\quad +584\cos(8u)+354\cos(9u)+164\cos(10u)+54\cos(11u)\\
&\quad +9\cos(12u))\sec(\tfrac{u}{2})^2\sec(u)^2\tan(\tfrac{u}{2}))/(1800u^2),
\end{aligned}$$

where $N = 57600u^2(1+\cos(2u)+\cos(4u))$. The local truncation error of this method is given by

$$LTER(h) = \frac{787}{12096}h^8(225\omega^6y^{(2)}(x)+259\omega^4y^{(4)}(x)+35\omega^2y^{(6)}(x)+y^{(8)}(x))+\mathcal{O}(h^{10}).$$

So this method is of order six, and we denote it as Method IV.

3 Numerical experiments

For the parameters $B = 0.002$ and $\omega = 1.01$ and for the initial values

$$y(0) = 0.200426728069666,\ y'(0) = 0,$$

Van Dooren [6] gave an accurately approximation to the analytic solution of Problem (1) as follows:

$$y(x) = \sum_{i=0}^{5} a_{2i+1}\cos((2i+1)\omega x),$$

where $A_1 = 0.2001794775366180, A_3 = 0.000246946143255837, A_5 = 3.04014985249\times10^{-7}$, $A_7 = 3.744\times10^{-10}, A_9 = 3.74349084\times10^{-10}, A_{11} = 5.68\times10^{-16}$. For comparison, we select a highly efficient two-frequency trigonometrically fitted Numerov method (denoted as TFNT) constructed by Fang et al. [7]. In our test, we choose the fitting frequencies $\omega_1 = \omega, \omega_2 = 3\omega$ for the method TFNT. The problem is integrated in the integral interval $[0, 1000]$ with several stepsize $h = 2/2^j, j = 0, 1, 2, 3$ and global errors are displayed in Table 1.

From Table 1 we can see that the new Methods II, III and IV are very effective for the integration of the undamped Duffing equation. Methods III and IV are more accurate than the method TFNT.

 ISBN: 978-84-608-6082-2

Table 1: The global errors at the right end point produced by the five methods for the undamped Duffing equation.

h	TFNT	Method I	Method II	Method III	Method IV
2	NaN	NaN	NaN	NaN	$2.9e-8$
1	0.15	NaN	NaN	NaN	$1.4e-7$
1/2	$8.7e-5$	$9.6e-4$	$1.7e-4$	$1.4e-5$	$1.1e-10$
1/4	$1.3e-6$	$8.6e-6$	$2.2e-6$	$8.5e-7$	$2.3e-10$
1/8	$1.9e-8$	$1.3e-7$	$3.1e-8$	$1.3e-9$	$1.6e-10$

4 Acknowledgements

This research is partially supported by NSFC (No. 11101357, 11571302) and the foundation of Scientific Research Project of Shangdong Universities (No. J14LI04).

References

[1] R.E. Mickens, *An Introduction to Nonlinear Oscillations,* Cambridge University Press, New York, 1981.

[2] Z. Wang, *Trigonometrically-fitted method with the Fourier frequency spectrum for undamped Duffing equation*, Comput. Phys. Commun. **174** (2006) 109-118.

[3] Z. Wang, *Obrechkoff one-step method fitted with Fourier spectrum for undamped Duffing equation*, Comput. Phys. Commun. **175** (2006) 692-699.

[4] T.E. Simos, *Exponentially-fitted and trigonometrically-fitted methods for long-term integration of orbital problems*, New Astron. **7** (2002) 1-7.

[5] Y. Fang, X.Y. Wu, *A trigonometrically fitted explicit Numerov-type method for second-order initial value problems with oscillating solutions*, Appl. Numer. Math. **58** (2008) 341-351.

[6] R. Van Dooren, *Stabilization of Cowells classical finite difference method for numerical integration*, J. Comput. Phys. **6** (1974) 186192.

[7] Y. Fang, Y. Song, X. Wu, *Trigonometrically fitted explicit Numerov-type method for periodic IVPs with two frequencies,* Comput. Phys. Commun. **179** (2008) 801-811.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Effective infection rate in SIR-type models from models with symptomatic and asymptomatic infection

Raquel Filipe[1], Nico Stollenwerk[1], Luís Mateus[1], Peyman Ghaffari[1], Scott Halstead[2] and Maíra Aguiar[1]

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

[2] *Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA*

emails: raquel.m.filipe@gmail.com, nico@ptmat.fc.ul.pt, luisgam1@yahoo.com, peyman@ptmat.fc.ul.pt, halsteads@erols.com, maira@ptmat.fc.ul.pt

Abstract

When studying some simple epidemiological models, such as the SIR model, for which we consider three types of individuals: susceptibles, infected or recovered. But what if one is infected and eventually infective but does not have any symptoms? How should such a dynamics be modeled in which we can have this kind of infected individuals? To do this we use a model which differentiates this type of infected individuals and so we have two types of infected: symptomatic and asymptomatic. We study the infection rate of an SIR model from this new model, hence an effective infection rate.

Key words: symptomatic, asymptomatic, effective infection rate

1 Introduction

Here we study an epidemiological model in which one can be susceptible, infected or recovered as in the simple SIR model. In our study we consider some not that often considered facts about the disease in study: we can have symptomatic and asymptomatic infected individuals. And so we divide the infected into two disjoint subsets. We supposed that the infected I are our empirical epidemiological data. And, so, we are considering a disease in which we can have four different kinds of individuals: susceptible S, notified and eventually hospitalized severe symptomatic H, asymptomatic or subsymptomatic and not notified infected A and recovered R. This model is denominated here by SHAR model.

  ISBN: 978-84-608-6082-2

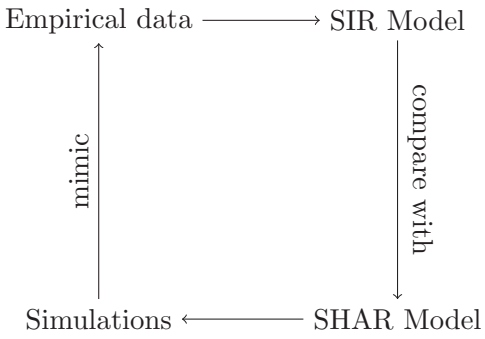
We assume our population size N to be constant, hence $N = S + H + A + R$. For the SHAR model we have the parameters $\alpha, \beta, \eta, \phi$ and γ, which we will describe in more detail later, and for SIR model we have the classical parameters $\tilde{\alpha}, \tilde{\beta}$ and $\tilde{\gamma}$. We will assume for the moment that $\tilde{\alpha} = \alpha$ and $\tilde{\gamma} = \gamma$. Our main objective is to study the effective infection rate $\tilde{\beta}$ as a function of the suposedly underlying $\alpha, \beta, \eta, \phi$ and γ. We introduce a relative contribution factor, ϕ, which represents how much more H contributes to the force of infection than A on meting susceptibles and infecting them. And we introduce also a rate η, which is given by the ratio of infected individuals who go to hospital over all infected individuals and so we have $\eta \in [0, 1]$. We will draw the rection schemes and the systems of differential equations and then we will computed the stationary states for the SHAR model and compare with those for the effective SIR model.

In empirical studies often not all details of the disease dynamics are known, but small compartmental models already can capture most if not all of the observed features of data series of notified disease cases. Simple models have easily higher probability than overcomplicated models given certain data, this in the sense of statistical model comparison via Bayes factors etc. [1]. Also it is not a priori clinically know if asymptomatically infected contribute in the same way as symptomatically and eventually severly infected persons, which might have a higher viral load than asymptomatic or subsymptomatic. This is especially important in vector borne diseases like e.g. dengue fever where a secondary infection with a second dengue virus strain can lead to an enhancement severity of disease due to an enhanced viral load. Then the question is if asymptomatically infected can transmit in the first place to mosquitoes in the same way as severely infected can. This question was recently investigated and answered positively for dengue fever specifically transmitted via the mosquitoes Aedes aegypti [2]. Finally, unlike in e.g. childhood diseases, which are highly infectious and symptomatic but humans developed effective immunity mechanisms, most diseases are less infective and also have often mild outcomes, so that we can expect some evolution towards decreasing pathogenicity in multi-strain populations, a mechanism coined "accidental pathogens". The basic mechanism is that severed disease is an evolutionary disadvantage for the underlying micro-organism which can grow best as commensal with the host population rather than harming hosts which then cannot transmit further the commensal turned pathogen. Finally, in populations with varying pathogenicity, the least pathogenic and hence mildly symptomatic or asymptomatic strains dominate the population [3, 4, 5, 6, 7]. Hence we would expect asymptomatically infected and infectious hosts to occur frequently rather than being the rare cases in epidemiology.

We model the SHAR complex situation with eventually many unknown classes and transition parameters in cases of multi-strain models, here in its simplest version and from this can generate via simulations data about the notified disease cases, here in its simplest version just the stationary state, which in turn can serve as input to simple effective models, like here the SIR model. Schematically, the research procedure sketched here can be given

 ISBN: 978-84-608-6082-2

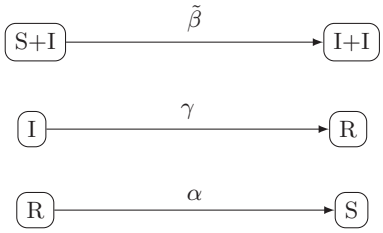
by



Finally, the simple effective models should be compared in performance on explaining the observed data with eventually necessary extensions, but only in case these extensions add explanatory value to the data analysis.

2 SIR Model

Schematically, the dynamics of the SIR model is given by



with infection rate $\tilde{\beta}$, recovery rate $\tilde{\gamma}$ and waining immunity rate, or in case of life long immunity a demographic renewal, $\tilde{\alpha}$ And so from the reaction scheme the system of differential equations is given by

$$\begin{cases} \dot{S} = \alpha R - \frac{\tilde{\beta}}{N} SI \\ \dot{I} = \frac{\tilde{\beta}}{N} SI - \gamma I \\ \dot{R} = \gamma I - \alpha R \end{cases}$$

with classically known stationary states given by

 ISBN: 978-84-608-6082-2

$$\begin{cases} S_1^* = N, I_1^* = 0, R_1^* = 0 \\ S_2^* = \frac{\gamma N}{\beta}, I_2^* = \frac{\alpha N}{\alpha+\gamma}\left(1-\frac{\gamma}{\beta}\right), R_2^* = \frac{\gamma N}{\alpha+\gamma}\left(1-\frac{\gamma}{\beta}\right) \end{cases}$$

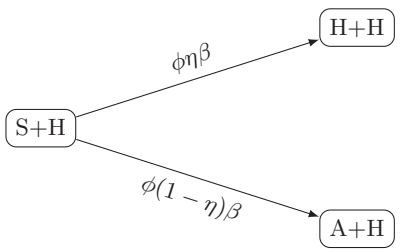
where (S_1^*, I_1^*, R_1^*) is the disease free stationary state and (S_2^*, I_2^*, R_2^*) the endemic stationary state, of further interest in the following.

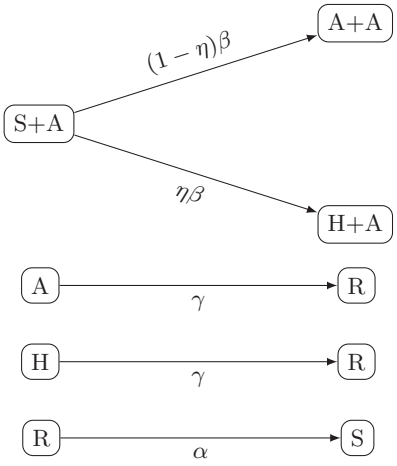
3 SHAR model

We now investigate an extension of the SIR model in the sense that we distinguish between infected being notified due to severe disease and eventually hospitalized, a disease class H, and infected not being notified due to mild or completely asymptomatic infection, a disease class A. Hence on infection of a susceptible S by meeting any of these infected H respectively A the notification ratio or hospitalization ratio η on top of the infection rate β sends the susceptible in the "hospital class" H, and with ratio $(1-\eta)$ into the "asymptomatic class". For $\eta = 1$ we wold come back to the old SIR model, since asymptomatic would not appear any more.

In case, hospital case H and asymptomatic case A can infect as susceptible in the same way, we already would have a complete model with only one new parameter. In case we want to consider the more general case of hospitalized cases H being more infective than asymptomatic case A, or eventually even less, e.g. due to immobilization in hospitale etc. as also [2] find as being possible in the case of dengue fever, we can include a ratio ϕ multipied with the infection rate β for hospitalized case meeting susceptibles. Then $\phi = 1$ is the above described simpler case, for $\phi > 1$ severe cases would transmit more the disease, and for $\phi < 1$ they transmit less. The limiting case of $\phi = 0$ gives a situation in which the asymptomatic cases trans mit, but not at all the severe hospitalized. But also the oposite case of hospitalized transmitting but not at all the asymptomatic can be captured in the present set-up, via the limit of ϕ becoming very large or even infinity while the infection rate β goes to zero in the same way, with no need for eventually two different parameters ϕ_H and ϕ_A, for hospitalized respectively asymptomatic.

Schematically, the dynamics of the SHAR model is gven by





and so the system of differential equations is given by

$$\begin{cases} \dot{S} = \alpha R - \frac{\beta}{N} S(A + \phi H) \\ \dot{H} = \eta \frac{\beta}{N} S(A + \phi H) - \gamma H \\ \dot{A} = (1-\eta) \frac{\beta}{N} S(A + \phi H) - \gamma A \\ \dot{R} = \gamma(A + H) - \alpha R \end{cases} .$$

For special cases of $\phi = 1$ or $\phi = 0$ the stationary states can be calculated easily along the lines of the calculations in the SIR model. But with some more effort also the general case for arbitray ϕ can be calulated giving the stationary states as follows

$$\begin{cases} S_1^* = N, H_1^* = A_1^* = R_1^* = 0 \\ S_2^* = \frac{\gamma N}{\beta(1+\eta(\phi-1))} \\ H_2^* = \eta \frac{\alpha}{\alpha+\gamma} \left(1 - \frac{\gamma}{\beta(1+\eta(\phi-1))}\right) N \\ A_2^* = (1-\eta) \frac{\alpha}{\alpha+\gamma} \left(1 - \frac{\gamma}{\beta(1+\eta(\phi-1))}\right) N \\ R_2^* = \frac{\gamma(H_2^* + A_2^*)}{\alpha} = \frac{\gamma}{\alpha+\gamma} \left(1 - \frac{\gamma}{\beta(1+\eta(\phi-1))}\right) N \end{cases} .$$

again with index 1 giving the disease free stationary state, and with index 2 the endemic stationary state.

4 Calculation of effective infection rate

The simplest way to estimate an effective infection rate $\tilde{\beta}$ of an SIR model from more complex models, like the here presented SHAR model, is to take the notified and severe

and eventually hospitalized cases H^* as the observed infected I^* of the effective SIR model, here of course in the endemic state Hence, we will calculate the effective infection rate, $\tilde{\beta}$, equaling H_2^* of the SHAR model to I_2^* of the SIR model, *i.e.* $I_{SIR}^* \stackrel{!}{=} H_{SHAR}^*$ and isolate $\tilde{\beta}$

$$\frac{\alpha\eta N(\beta\phi\eta - \beta\eta + \beta - \gamma)}{\beta(\alpha+\gamma)(\phi\eta - \eta + 1)} = \frac{\alpha N}{\alpha+\gamma}\left(1 - \frac{\gamma}{\tilde{\beta}}\right)$$

$$\Rightarrow \tilde{\beta} = \frac{\gamma}{1 - \eta\left(1 - \frac{1}{1+\eta(\phi-1)}\frac{\gamma}{\beta}\right)}.$$

Now for various values of ϕ we can plotted the effective infection rate as a function of the notification or hospitalization ratio η by fixing the other parameters, see Fig. 1.

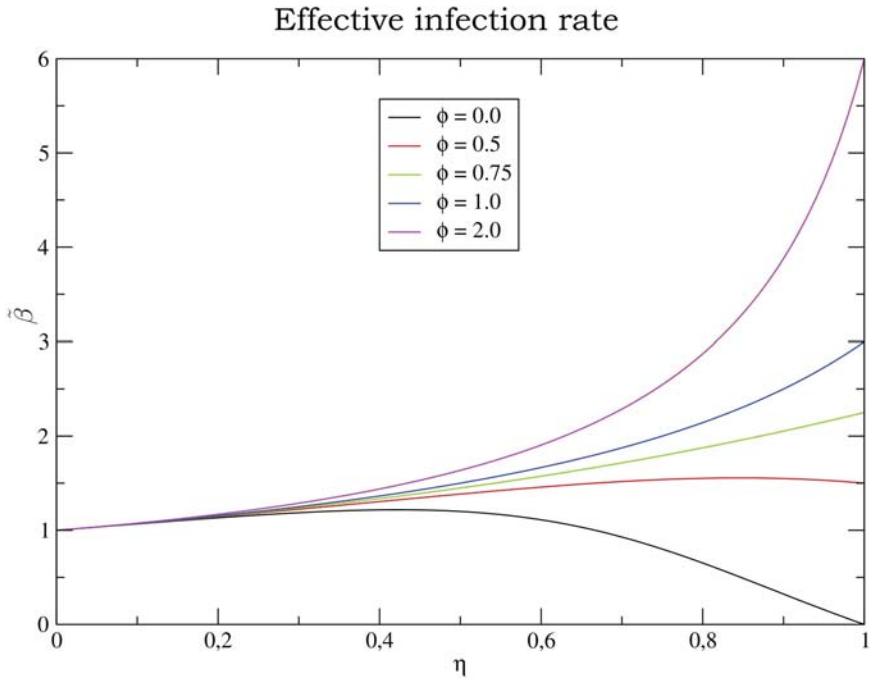


Figure 1: Effective infection rate, $\tilde{\beta}$, for different values of ϕ

For the present theoretical study we fix the parameters of the SHAR model to $\gamma = 1$ hence the time units are fixed to the recovery period and the infection rate to $\beta = 3 \cdot \gamma$, while α is not appearing in the analytic result of the effective infection rate $\tilde{\beta}$. Hence in Fig. 1 we find for $\phi = 1$ and $\eta = 1$ the value of $\tilde{\beta} = \beta = 3$. And again for $\phi = 1$ and now for e.g. $\eta = 0.75$ we obtain from an underlying $\beta = 3 \cdot \gamma$ an effective $\tilde{\beta} = 2 \cdot \gamma$.

In initial studies we already also found in time dependent solutions $H(t)$ versus $I(t)$ that the presently calculated $\tilde{\beta}$ gives a good comparison not only i stationarity but also in transients. Such results could eventually be improved by also relaxing the assumptions of $\tilde{\alpha} = \alpha$ and $\tilde{\gamma} = \gamma$ (and later concerning stochastic fluctuations the case of an effective system size $\tilde{N}$ in the SIR model with $\tilde{N} \neq N$ could be investigated).

5 Conclusions and future research

We calculated an effective infection rate $\tilde{\beta}$ of a simple SIR model depending on further parameters like notification rate η and contribution to the force of infection of notified infected versus asymptomatic ϕ, resulting in mostly slight changes of $\tilde{\beta}$ versus the underlying infection rate β of more complex models (respectively $\phi \cdot \beta$ for varying ϕ). In future studies stochastic realizations of the here described SHAR model could be used to estimate the effective parameters of a simple SIR model (now with free parameters not only the infection rate $\tilde{\beta}$ but also other parameters $\tilde{\alpha}$, $\tilde{\gamma}$ and $\tilde{N}$), and this not only for fluctuations around the endemic stationary state but also in transients. In certain parameter regions and for few data points, hence littel information of the complexity of the underlying model, here the SHAR model, we would again expect in Bayesian model comparison higher probability for simpler models like an effective SIR model, as we found in simpler epidemiological models which still could be treated analytically [1]. Now more numerical methods like described e.g. in [8] would come to place. Especially in studies of multi-strain models describing dengue fever it turned out that primary versus seondary infection is more important in describing the observed fluctuations in empirical data e.g. from dengue hospital notification data from Thailand than the exact number of strains [9], not needing all possible model complexities. Of course a good understanding of the notification data is of importance [10].

Acknowledgements

This work has been supported by the European Union under FP7 in the project DENFREE and by FCT, Portugal, in various ways.

References

[1] Mateus, L., Stollenwerk, N., & Zambrini, J.C. (2013) Stochastic Models in Population Biology: From Dynamic Noise to Bayesian Description and Model Comparison for Given Data Sets, *Int. Journal. Computer Math.* **90**, 2161–2173.

[2] Veasna Duong, Louis Lambrechts, Richard E. Paul, Sowath Ly, Rath Srey Laya, Kanya C. Long, Rekol Huy, Arnaud Tarantola, Thomas W. Scott, Anavaj Sakuntabhai, and

Philippe Buchy (2015) Asymptomatic humans transmit dengue virus to mosquitoes, *Proc. Nat. Acad. Science* **112**, 14688–14693.

[3] Stollenwerk, N., & Jansen, V.A.A. (2003, a) Meningitis, pathogenicity near criticality: the epidemiology of meningococcal disease as a model for accidental pathogens. *Journal of Theoretical Biology* **222**, 347–359.

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

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

[6] N. Stollenwerk and V. Jansen (2011) *Population Biology and Criticality: From critical birth–death processes to self-organized criticality in mutation pathogen systems* (Imperial Colledge Press, World Scientific, London).

[7] Peyman Ghaffari, Vincent Jansen & Nico Stollenwerk (2011) Evolution towards critical fluctuations in a system of accidental pathogens, ICNAAM 2011, Chalkidiki, 1263–1266. (Numerical Analysis and Applied Mathematics ICNAAM 2011 AIP Conf. Proc. 1389, 1224-1227 (2011); doi: 10.1063/1.3637837 Copyright 2011 American Institute of Physics 978-0-7354-0956-9)

[8] Stollenwerk, N., Aguiar, M., Ballesteros, S., Boto, J., Kooi, W. B., & Mateus, L. (2012). Dynamic noise, chaos and parameter estimation in population biology, *Interface, Focus*,**2**, 156–169.

[9] Aguiar, M., Kooi, W. B., Rocha, F., Gaffari, P. & Stollenwerk, N. (2013). How much complexity is needed to describe the fluctuations observed in dengue hemorrhagic fever incidence data? *Ecological Complexity*, **16**, 31–40.

[10] Aguiar, M., Paul, R., Sakuntabhai, A., & Stollenwerk, N. (2014) Are we modeling the correct data set? Minimizing false predictions for dengue fever in Thailand, *Epidemiology and Infection*, **142**, 2447–2459.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Lyapunov spectra for torus bifurcations and ways to deterministic chaos in population biology

Pablo Fuentes Sommer[1], Nico Stollenwerk[1], Bob Kooi[2], Luís Mateus[1], Peyman Ghaffari[1] and Maíra Aguiar[1]

[1] *Centro de Matemática, Aplicações Fundamentais e Investigação Operacional, Universidade de Lisboa, Faculty of Science, Campo Grande, Lisboa, Portugal*

[2] *Department of Earth and Life Sciences, Vrije Universiteit, Amsterdam, The Netherlands*

emails: p_fuentes@outlook.com, nico@ptmat.fc.ul.pt, bob.kooi@vu.nl, lgmateu@fc.ul.pt, pgsaid@fc.ul.pt, mafsantos@fc.ul.pt

Abstract

A simple ecological population system like the Rosenzweig-MacArthur model with a Holling Type II response function illustrates a Hopf bifurcation. If we add a seasonal variation to one of the parameters the Hopf bifurcation point becomes a torus bifurcation and further through torus destruction we get deterministically chaotic behavior. We study here the Lyapunov spectra to get insight about the the locations of torus bifurcations, torus destruction and chaotic regions in parameter space. Especially interesting are detected coexisting attract ors, characterized by different Lyapunov exponents depending on initial conditions of the trajectories on which the Lyapunov exponents are calculated. As the trajectories converge to different dynamic attractors the Lyapunov exponents change, and sometimes abruptly when constant initial conditions are used.

Key words: Rosenzweig-MacArthur model, torus bifurcation, deterministic chaos, Lyapunov spectra, coexisting attractors.

1 Introduction

One of the simplest models to study Hopf bifurcations and hence with seasonal forcing also torus bifurcations is the Rosenzweig-MacArthur model, which however has no direct stochastic interpretation due to its time scale separation leading to the Holling type II response function. We studied this system initially in [1] in respect to the time scale separation, and stochastic versions via time continuous and state discrete Markov processes and

Fokker-Planck approximations, see also [2]. The seasonally forced Rosenzweig-MacArthur model has long been investigated [3], finding torus bifurcations and period doublings into chaotic motion, where however due to the used methods of bifurcation software only fixed points and limit cycles can be tracked, but no information can be obtained beyond the torus bifurcation, hence leaving some insecurity as to where the chaotic attractors should appear. We overcome this difficulty by analyzing Lyapunov spectra, which can detect deterministically chaotic attractors via positive Lyapunov exponents, but also can indicate bifurcations since at period bifurcations one Lyapunov exponent becomes zero and is negative on both sides of the bifurcation. Tori are characterized by two largest zero Lyapunov exponents.

To study stochastic versions of torus bifurcations and torus break up into chaos, the parameter values in [3] are not suitable, since it is not easily seen how a time scale separable system can be obtained from this. But finally we found suitable parameter values which have similar bifurcation structures as found in [3], but also allow the time scale separation, hence stochastic versions as investigated e.g. in [1]. On the way to detect the torus bifurcations we found, like [3] with their bifurcation software, with our methods of Lyapunov spectra coexistences of attractors in such a relatively simple ecological system, and only via suitable initial conditions we could detect the regions of torus bifurcations. These findings have, of course, wide implications in other population biological systems, in which torus bifurcations and soon after deterministic chaos have been detected [4]. In that case also bifurcation software like AUTO was used to initially find the torus bifurcations. The stochastic versions of such systems are of importance in further theoretical understanding [5, 6, 7] and also empirical data analysis [8].

2 The Rosenzweig-MacArthur model

We consider the Rosenzweig-MacArthur model studied in [1] and [2]. In this model X represents a population of preys that consume resources S; a population of predators Y feeds on the preys and they can be in one of two disjoint classes, Y_s represents a "searching" class that is looking for prey and Y_h is a "handling" class which has already fed on some prey and will therefore not be searching for more. The model is described by the stoichiometric formulation

$$S + X \xrightarrow{\beta} X + X \tag{1}$$
$$X \xrightarrow{\alpha} S \tag{2}$$
$$X + Y_s \xrightarrow{b} Y_h \tag{3}$$
$$Y_h \xrightarrow{k} Y_s \tag{4}$$

$$Y_h \xrightarrow{\nu} Y_h + Y_s \tag{5}$$
$$Y_s \xrightarrow{\mu} \emptyset \tag{6}$$
$$Y_h \xrightarrow{\mu} \emptyset \tag{7}$$

Reaction (1) represents a prey consuming resources to procreate at a rate β. Reaction (2) shows preys dying at a rate α and giving resources free. In reaction (3) we model searching predators who hunt preys and change their state to handling predators at a rate b. Handling predators become hungry again and change their state to searching at a rate k in reaction (4). Handling predators can also give birth to other predators at a rate ν, these newborn predators are assumed to be in searching state in reaction (5). Reaction (6) and (7) model predators dying at a rate μ.
We assume that the maximal prey population $N := X + S$ is constant. We also define the growth rate of the preys ϱ and the carrying capacity κ [6]

$$\varrho := \beta - \alpha \qquad \text{and} \qquad \kappa := N\left(1 - \frac{\alpha}{\beta}\right) \quad ,$$

and we get a three dimensional ordinary differential equation system:

$$\dot{X} = \varrho X\left(1 - \frac{X}{\kappa}\right) - \frac{b}{N}XY_s \tag{8}$$
$$\dot{Y_h} = \frac{b}{N}XY_s - kY_h - \mu Y_h \tag{9}$$
$$\dot{Y_s} = -\frac{b}{N}XY_s + kY_h - \mu Y_s + \nu Y_h \quad . \tag{10}$$

2.1 Time scale separation

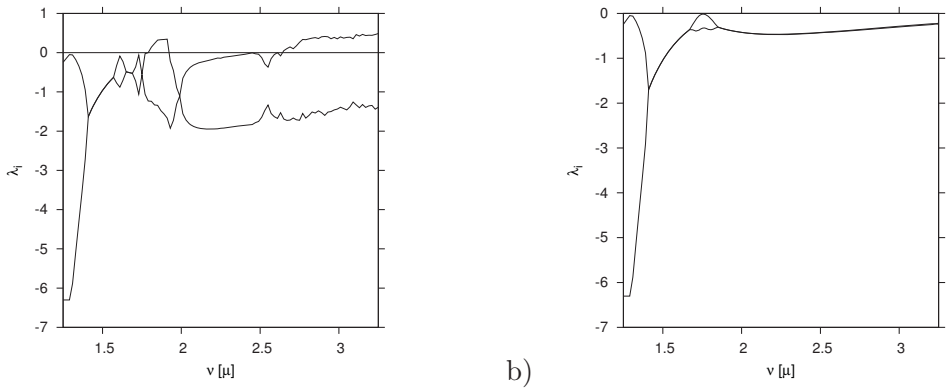
Using a time scale separation argument as described in [1] we get a two dimensional ordinary differential equations system for a Rosenzweig-MacArthur model with Holling type II response function

$$\dot{X} = \varrho X\left(1 - \frac{X}{\kappa}\right) - k\frac{X}{\frac{k}{b}N + X}Y$$
$$\dot{Y} = -\mu Y + \nu\frac{X}{\frac{k}{b}N + X}Y \quad .$$

For the model to fulfill a time scale separation we asign values to the parameters that conform an ecological model as in [9]. In this sense the preys have a lifespan of roughly

 ISBN: 978-84-608-6082-2

half a year, so $\alpha := \frac{1}{\frac{1y}{2}} = 2y^{-1}$. Assuming they procreate fast gives $\beta := 20\alpha$. We assume that the predators have a lifespan of 10 months and a procreation rate twice as fast as the mortality, so $\mu := \frac{1}{\frac{10y}{12}} = \frac{6}{5}y^{-1}$ and $\nu := 2\mu$. Further, the predators would have a digestion rate of 1 day, hence $k := \frac{1}{1d} = 365^{-1}$. And finally a hunting rate $b := 3k$.



a) b)

Figure 1: *Lyapunov spectra with $t_{max} = 500$ for seasonal forcing a) $\eta = 0.5$ with clearly positive Lyapunov exponents for some parameter regions and b) $\eta = 0.1$ with Lyapunov exponents occasionally close to zero, indicating a torus as attractor. Parameters as given in [3].*

3 Seasonality and Torus

3.1 Seasonal forcing

Past studies have considered seasonal variation of different parameters [3]. We follow [9] determining the predator hunting rate as the most suitable parameter for seasonal variation. Yielding a seasonaly forced system

$$\begin{aligned} \dot{X} &= \varrho X\left(1-\frac{X}{\kappa}\right) - k\frac{X}{\frac{k}{b}N+X}Y \\ \dot{Y} &= -\mu Y + \nu(t)\frac{X}{\frac{k}{b}N+X}Y. \end{aligned}$$

with

$$\nu(t) = \nu_0(1+\eta cos(\omega(t+\phi))) \quad .$$

Then we can couple a Hopf oscillator [9] to get an autonomous systme for which we can calculate the Lyapunov exponents [5, 9]. The autonomous system coupled with the Hopf

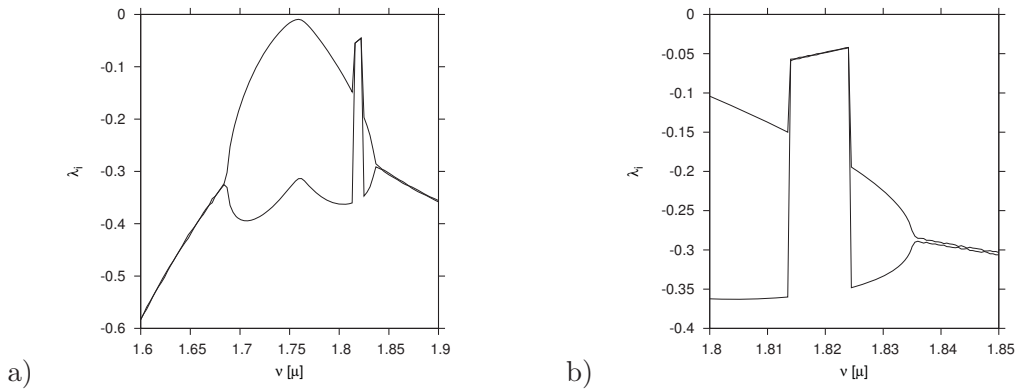


Figure 2: *Lyapunov spectra with* $t_{max} = 500$ *for seasonal forcing* $\eta = 0.1$*. Zooming into the parameter space reveals co-existing attractors around the region of the torus bifurcation. a) and b) for fixed initial values* $X(t_0) = 0.167$ *and* $Y(t_0) = 0.0015$ *for each value of mean birth rate of predators* ν_0*. Parameters as given in [3].*

oscillator shows to have the same Lyapunov exponents as the forced system apart from the Lyapunov exponents of the oscillator itself [9]. Hence we can directly use the forced system to calculate the two Lyapunov exponents which are non-trivial, baring in mind that the trivial zero Lyapunov exponent along a trajectory is in addition to take into account (as well a second trivial exponent results from the contraction along the Hopf oscillator).

3.2 Lyapunov exponents of the forced Rosenzweig-MacArthur model

Now we examine closely the behaviour of the Lyapunov exponents for variable ν_0 in order to get an insight on the chaotic behaviour of the system. Thus, for fixed initial conditions we expect one Lyapunov exponent to be zero at the Hopf bifurcation point, whenever an exponent is positive it indicates deterministic chaos and for a non-trivial exponent close to zero (besides the trivial zero Lyapunov exponent) we expect there to be a torus as an attractor, see Figure 1.

Figure 1 b) shows what could be interpreted as a torus bifurcation at around $\nu_0 = 1.8\mu$. But zooming in shows a jump in the otherwise relatively smooth curves of the Lyapunov exponents, see Figure 2. These jumps on the values of the Lyapunovs exponents turn out to be jumps on the trajectories between coexisting attractors. They can be avoided by changing the initial conditions in order to track one specific attractor, which yields a rather smooth curve for the Lyapunov exponents, see Fig. 3 with a) showing one attractor and b) showing the detection of the torus bifurcation by starting in the jump region with $\nu_0 = 1.82\mu$.

Tracking of the torus gives values of the state variables on the torus, here for $\nu = 1.86\mu$

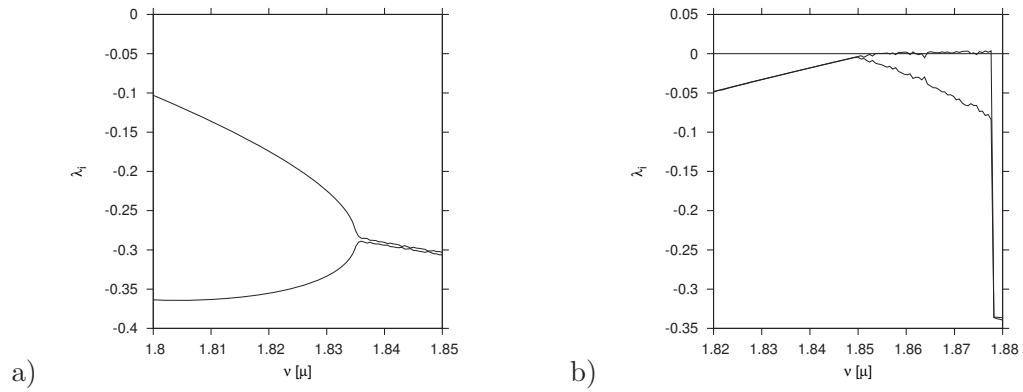


Figure 3: *Tracking the different attractors from the initial values with parameter* ν_0.

as $X = 0.380010$ and $Y = 0.162432$, which now can be used as new fixed initial values $X(t_0) = 0.380010$ and $Y(t_0) = 0.162432$, giving a Lyapunov spectrum, mainly in the torus region. The Lyapunov spectrum given with fixed initial values $X(t_0) = 0.380010$ and $Y(t_0) = 0.162432$ is shown in Fig. 4.

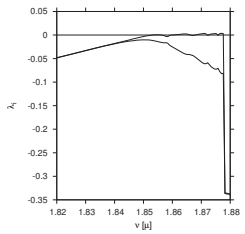


Figure 4: *Lyapunov spectra with* $t_{max} = 500$ *for seasonal forcing* $\eta = 0.1$*, for fixed initial values now on the torus* $X(t_0) = 0.380010$ *and* $Y(t_0) = 0.162432$*. Parameters as given in [3].*

3.3 Lyapunov characteristic exponents for two varying parameters

After we have detected the torus bifurcation, see Figure 4, we now can use these intial conditions to vary two parameters at the same time, the mean predator birth rate ν_0 and the seasonality η and plot in color the leading Lyapunov exponent, also called Lyapunov characteristic exponent.

Since we have used the forced Rosenzweig-MacArthur 2-dimensional system, we could eliminate elegantly the trivial zero Lyapunov exponent. In Figure 5 a) hence a negative characteristic exponent, in blue or green, indicates in the forced system a limit cycle, which

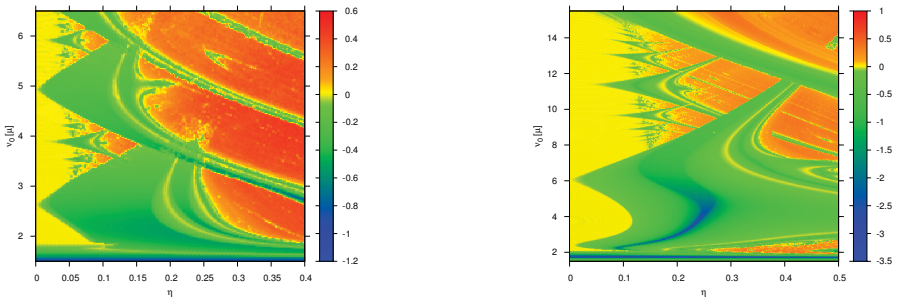


Figure 5: *Lyapunov characteristic exponents for two varying parameters, the mean predator birth rate ν_0 and the seasonality η, a) in good comparison to what [3] have found with different techniques like bifurcation software, and b) for our parameter values allowing time scale separation.*

in the unforced system, hence $\eta = 0$ would be a fixed point. Then the zero characteristic exponent indicates the region of torus dynamics, in yellow. In the yellow regions we find again limit cycles cutting in towards vanishing seasonality, the Arnol'd tongues, which widen for increasing seasonality η.

And only in these Arnol'd tongues we find characteristic exponents indicating bifurcations into chaotic dynamics, i.e. positive Lyapunov exponents in orange and red, hence the phase locking on the tori breaks up into more than 2 dimensional dynamic attractors, since three dimensions are needed for deterministic chaos to allow non-returning attractors different from limit cycles.

Finally in Figure 5 b) we find the same dynamical behaviour also for our parameter values which allow a time scale separation from a more complex stoichiometric system to the two-dimensional Rosenzweig-MacArthur model, as analyzed previously in [1]. This allows of course also in future studies the analysis of stochastic systems derived from the stoichiometric system [6]. So we can analyze the stochastic behaviour around the torus bifurcation and torus break up into deterministic chaos. This is of major interest especially in analyzing real world systems, like dengue fever epidemiology with torus bifurcations [4] and empirical data to perform parameter estimation, see [5, 7, 8].

Acknowledgements

This work has been supported by the European Union under FP7 in the project DENFREE and by FCT, Portugal, including an FCT-DAAD exchange grant.

References

[1] Pablo Fuentes Sommer, Luís Mateus, Bob Kooi, Maíra Aguiar and Nico Stollenwerk (2015) Hopf and torus bifurcations in stochastic systems in mathematical population biology. *Proceedings of the 15th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2015, Cadiz, Spain*, pp. 543–555, ISBN: 978-84-617-2230-3, edited by Jesus Vigo et al.

[2] Nico Stollenwerk, Pablo Fuentes Sommer, Luís Mateus, Bob Kooi & Maíra Aguiar (2015) Stochastic Hopf and torus bifurcations in population biology, *ICNAAM, Numerical Analysis and Applied Mathematics, International Conference 2015, Rhodes, Greece*, edited by T.E. Simos, G. Psihoyios, and Ch. Tsitouras, American Institute of Physics.

[3] Rinaldi, S., Muratori, S., & Kuznetsov, Y (1993) Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities, *Bulletin of Mathematical Biology* **55**, 15–35.

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

[5] Aguiar, M., Ballesteros, S., Kooi, B.W., & Stollenwerk, N. (2011) The role of seasonality and import in a minimalistic multi-strain dengue model capturing differences between primary and secondary infections: complex dynamics and its implications for data analysis, *Journal of Theoretical Biology*, **289**, 181–196.

[6] Stollenwerk, N., & Jansen, V. (2011) *Population Biology and Criticality: From critical birth–death processes to self-organized criticality in mutation pathogen systems* (Imperial College Press, World Scientific, London).

[7] Aguiar, M., Stollenwerk, N. & Kooi, W. B. (2012). Scaling of stochasticity in dengue hemorrhagic fever epidemics. *Math. Model. Nat. Phenom.*, **7**, 1–11.

[8] Stollenwerk, N., Aguiar, M., Ballesteros, S., Boto, J., Kooi, W. B., & Mateus, L. (2012). Dynamic noise, chaos and parameter estimation in population biology, *Interface, Focus*,**2**, 156–169.

[9] Nico Stollenwerk, Pablo Fuentes Sommer, Bob Kooi, Luís Mateus, Peyman Ghaffari & Maíra Aguiar (2016) Hopf and torus bifurcations, torus destruction and chaos in population biology, *submitted for publication.*

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Black-List Genetic Algorithm Scheduling for Energy Saving in Heterogenous Environments

Eloi Gabaldon[1], Fernando Guirado[1], Josep Lluis Lerida[1] and Jordi Planes[1]

[1] *Department of Computer Science, Universitat de Lleida*

emails: eloigabal@diei.udl.cat, f.guirado@diei.udl.cat, jlerida@diei.udl.cat, jplanes@diei.udl.cat

Abstract

Reducing energy consumption in large scale computing facilities has become a major concern in recent years. Most of the techniques have been focused on determining the computing requirements based on load predictions and thus turning unnecessary nodes on and off. Nevertheless, once the available resources have been configured, new opportunities arise for reducting energy consumption by providing optimal matching of parallel applications to the available computing nodes. Those techniques have not received much attention otherwise. The large number of computing nodes, heterogeneity and application-tasks variability are factors that turn the scheduling into an NP-Hard problem. Proposed in this paper is a genetic algorithm based on a weighted blacklist able to generate scheduling decisions that globally minimize the energy consumption.

Key words: Weighted Blacklist, Energy Saving, Genetic Algorithm, Federated Clusters, Scheduling, Co-allocation

1 Introduction

The computing requirements of scientific applications are continuously growing, as is the amount of data to process. The use of more sophisticated and scalable infrastructure becomes necessary to cover these requirements. Different architectural environments such as Federated Clusters, Grid infrastructures or Cloud computing are examples of federated resource environments, where computing resources in different administrative domains work together to solve a problem [1, 2].

Nevertheless, the high energy consumption produced by the use of large-scale resources translates into high energy cost and high carbon emissions which are not environmentally sustainable. Hence, there is an urgent need for energy-efficient solutions that force s redefinition of the way computing resources must be used. Recent research is able to determine the quantity of resources to be used based on load predictions, QoS requirements, etc. Nonetheless, once selected the required resources, new opportunities for reducting energy consumption arise from providing optimal matching of parallel applications to the available computing nodes. These techniques otherwise have otherwise received little attention.

In this work, we focus on Federated Cluster environments. These environments allow the execution of parallel applications in which the computing resource requirements exceed the resources available in a single cluster. Thus, some parallel applications can be co-allocated to different clusters, sharing computational and communication resources. Under these circumstances, a large amount of resources, their heterogeneity and also the ability to apply co-allocation of tasks among different administrative domains turns the job scheduling problem into an NP-hard problem. However, the adequate matching of computational nodes and parallel tasks can provide an improvement to the final energy saving process.

The main goal in the present paper is to provide a technique that considers the computing heterogeneity, the parallel tasks requirements and energy resource consumption to generate scheduling decisions that minimize the energy consumption of the workloads to be executed. The contribution of this paper is a novel scheduling technique based a genetic algorithm guided by a weighted blacklist of forbidden resources. The results obtained from scheduling different workloads from real traces shows that our proposal is able to reduce the energy consumption compared with other techniques from the literature.

The remainder of this paper is organized as follows. Section 2 presents related work. The proposed genetic algorithm is elaborated in Section 3. Section 4 demonstrates the performance analysis and the simulation results for real workload traces. Finally, the conclusions is presented in Section 5.

2 Related Work

The potential benefit of sharing computing resources among independent sites in federated environments has been widely discussed in previous research [1, 2]. However, resource heterogeneity, data transferring and contention in the communication-links have a large influence on the execution cost of parallel applications, and become critical aspects for the exploiting resources and application performance [3, 4].

Beside this, the benefits of aggregating resources to solve huge computational problems becomes a critical issue from the point of view of energy saving. Some research in the literature has been conducted with the aim of reducing energy consumption. Several studies are focused in determine the energy consumption of a computing system. Srikantaiah et

al. [5] show that the energy consumed by a machine depends directly on the CPU usage. Furthermore, Chia-Hung Lien et al. [6] demonstrate that the consumption of a modern CPU in the idle state can be over 50% of its consumption at maximum power. Thus, the best approach corresponds to using only those computing resources that are really necessary to execute the workload and turning the others off.

A large amount of research was previously conducted to predict the future load in order to turn the resources on and off depending on the computing requirements. Cocaña et al. [7] presented a software tool that predicts the future node requirements using a machine-learning approach, and then stopping those that will not be required in the near future. In a similar way, Chae et al. [8] illustrate a method of determining the aggregate system load and the minimal set of computational resources that can process the workload. Pinheiro et al. in [9] deal with turning the computational resources on-off in order to handle with a given workload. And Orgerie et al. in [10] present a three-step strategy based on a framework able to control the computing requirements by switching the unused nodes off, predicting their usage in order to switch on them on again and finally aggregating some reservations in order to avoid frequent on/off cycles.

Once the required computational nodes are identified and configured, the scheduling process must allocate the tasks of the parallel applications to them. Scheduling decisions at this stage are critical because energy consumption depends directly on the policy used. Nonetheless, less attention has been paid to this field of research. Li et al. [11] presented a batch technique using a Min-Min heuristic in which the computing nodes can be in different consumption states. When the have not been used nodes are not being used for a period of time, they change to a lower consumption state using a OPSCS strategy [12]. Hsu et al. [13] identifies an energy consumption threshold based on the resource utilization and allocates the parallel applications to avoid exceeding this value.

In the present paper, the main contribution is the implementation of a weighted blacklist that determines the resource unavailability for each parallel job to be scheduled. Our proposal avoids any systematic allocation of the jobs to the resources with minimum consumption, as is usual in the literature, and it provides new scheduling chances for the ramaining jobs in the system queue. Our proposal is integrated into a Genetic Algorithm based meta-heuristic, named GA-BL for Genetic Algorithm BlackList, which is able to modify the job execution order to take profit of the free resources from the blacklist.

3 Blacklist Energy-Aware Genetic Algorithm

A Genetic Algorithm (GA) is a search heuristic to find near-optimal solutions using nature-based techniques. It starts by creating an initial population of solutions known as individuals, each one encoded using a chromosome. To create a new generation, four steps are performed: ranking the individuals driven by a fitness function, ranking-based selec-

tion, crossover and mutation. The algorithm is motivated by the hope that after several generations, the new population will be better than the older ones.

In this paper, the authors propose the integration of a weighted blacklist of the computational nodes, which represents the nodes forbidden to be used in the job allocation process, into a GA meta-heuristic. This novel *GA-BL* is mainly focused on reducing energy consumption. This energy consumption is modeled with equation 1. Considering that, in a real system, the computing nodes (N) remain in one of two states, *idle* or *computing*. Each node consumes a different amount of energy depending on the current state, with nodes that are computing consuming more energy than the ones that are idle. Energy consumption is calculated as follows:

$$\sum_{n}(E_n^c * T_n^c + E_n^i * T_n^i) \quad \forall n \in \ N \tag{1}$$

E_n^s being the energy consumed by node n when it is in state s, where i stands for *idle* and c stands for *computing*, and T_n^s is the time spent by the node n in the state s. The execution time of the job is calculated using the model presented by the authors in [14].

The first decision in developing a GA is the chromosome design. This is used to represent the individuals in the population. The chromosome representation used is composed of two parts. The first part represents the order the jobs are to be executed in the system. The second part represents the weighted blacklist resource unavailability for each job (see Figure 1). The blacklist was implemented as a real number in the range $[0, 1]$ for each job and cluster, the value represents the percentage of cluster nodes forbidden from the allocation of the corresponding job.

To obtain the final job scheduling, the heuristic described in Algorithm 1 was used. If, in the allocation process, we run out of computational nodes, GA predicts the first job to finish by using the job execution model presented in [14] and then it releases the allocated nodes for the subsequent jobs.

Algorithm 1 Allocation Algorithm

Require: $\mathcal{Q}$: Set of jobs
Ensure: $\mathcal{A}$: Set of $(Task, Node)$
1: **for** $Job \in \mathcal{Q}$ **do**
2: $N \leftarrow FreeNodes - ForbiddenNodes$
3: **for** $Task \in Job$ **do**
4: $\mathcal{A} \leftarrow \mathcal{A} \cup (Task, \text{argmin}_{n \in N}(Energy(n)))$
5: **end for**
6: **end for**

Example 1. Given a set of jobs $\mathcal{J} = \{J1, J2, J3, J4, J5\}$, having every job a set of tasks as follows: $J1 = \{T1, T2, T3, T4\}$, $J2 = \{T5\}$, $J3 = \{T6, T7, T8\}$, $J4 = \{T9, T10, T11, T12,$

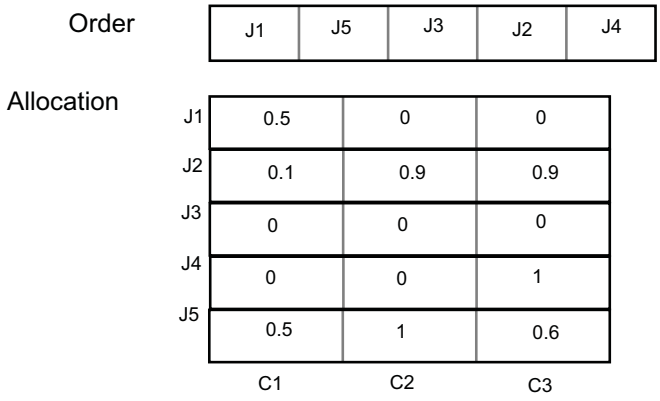


Figure 1: GA representation for Example 1

$T13\}$ and $J5 = \{T14\}$; and a set of clusters $\mathcal{C} = \{C1, C2, C3\}$, every cluster has a set of nodes as follows: $C1 = \{N1, N2\}$, $C2 = \{N3, N4, N5\}$, and $C3 = \{N6, N7\}$; a possible solution is given in Figure 1. Observe that job J4 is the last job to be scheduled and it cannot use any of the nodes in C3, since the solution has a value of 1.0 in its node allocation spot, what means that 100% of the nodes in this cluster are forbidden for this job. The corresponding scheduling is represented in Figure 2. For the sake of clarity, the job execution time and energy consumption for all machines are considered the same. Note that J4 has as forbidden nodes $\{N5, N6\}$, so it will wait for N1 to be scheduled.

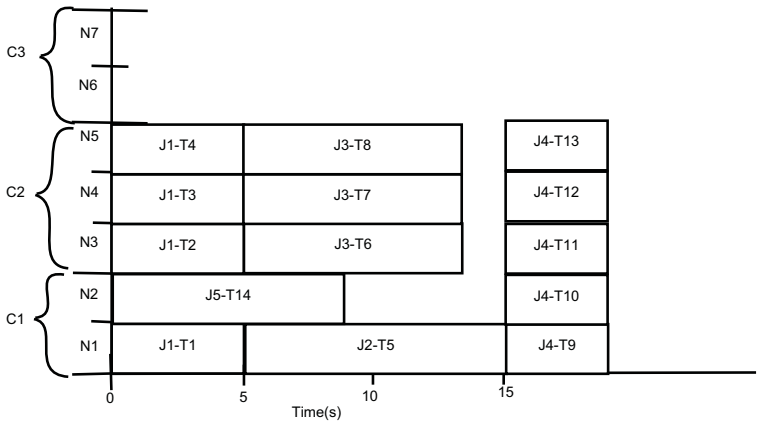


Figure 2: Schedule for Example 1

To evaluate each individual in the population, the GA-BL uses the **fitness function** described in Equation 1. In each iteration, the individuals with better fitness are selected using a standard tournament selection algorithm. The **crossover** operator is divided into two different algorithms to deal with the representation.

Order: a mask of random binary values is generated. For every position with value 1, the job of the first parent is placed in the offspring. For the missing jobs, the order of the second parent is chosen.

Allocation: we use a real number crossover for each value in the allocation grid. A random value α is selected between the interval[-0.2,1.2] to avoid the premature convergence of the algorithm. Then, the result of the crossover is calculated with the equation $F_{result} = \alpha * F_{parent1} + (1 - \alpha) * F_{parent2}$.

The **mutation** operator used to maintain the diversity through the generations is also divided into two parts.

Order: this consist of swapping two jobs in the allocation queue

Allocation: a new randomly-selected value is obtained.

4 Experimentation

In this section we conducted an experimental study with the aim of analyzing the reduction in energy consumption obtained by applying the proposed GA-BL scheduling technique. The experimentation was carried out by simulation, using the Gridsim simulator configured to emulate an environment composed of 4 different clusters. The characteristics of the clusters are shown in Table 1. The consumptions of the different nodes were estimated by using the study done in [10] where the consumption of an idle node and a computing one are identified.

Cluster	Nodes	Effective Power MIPS	Energy idle Kw/s	Energy Computing KW/s
1	64	1000	130	300
2	64	1200	150	320
3	64	1300	250	400
4	64	1800	280	400

Table 1: System characteristics.

The evaluation process has been carried out with other well known techniques existing in the literature able to be used on Federated Cluster environments.

These techniques were *JPR-E*, a variant of Naik's heuristic [15], where the tasks are matched with the less energy consuming nodes and *FCFS*, a technique that schedules the first job in the queue to available nodes. Both techniques treats only individual jobs from the queue. *Min-Min* is a heuristic based on [11]. It is able to consider a set of jobs with the aim of minimizing the energy consumption by the nodes. And Finally, *HILL* is a well-known hill-climbing meta-heuristic focused on minimizing the fitness function described in Equation 1.

The set of parameters that guide our genetic algorithm proposal is shown in Table 2. These parameters were selected as a trade-off between the improvement in performance and the computational cost according to the experimentation done in [14].

Parameter	Value
Num. Iterations	120
Population Size	80
Mutation Frequency	50

Table 2: GA-BL parameters.

We evaluated three different workloads HPC2N, CEA CURIE and RICC extracted from real cluster traces described by Feitelson [16].

Half of the jobs in the *HPC2N* are made up of 1-2 tasks and the rest, evenly distributed in 4,8,16 tasks. The average execution time of this workload is around 1 hour.

80% of jobs in the *RICC* workload are made of 1 tasks while the rest can be up to 32, and the execution times of each job can vary greatly between 10 seconds to 10 hours.

In the *CEA CURIE* workload, approximately 50% of the jobs are made up of 32 tasks, with the execution time of 80% of the jobs being extremely low, from 1 second to 1 minute.

Several job-sets corresponding to one month of execution time were selected for each workload. To evaluate the energy consumption of all sets, we used a boxplot that synthesizes the most relevant information of the results: minimum, median and maximum consumption values as well as the frequency of jobs in the first and third quartile. The outliers are displayed as dots.

Figure 3 shows the results obtained for the HPC2N workload. As we can see, GA-BL gave the lowest consumption with lower minimum and maximum values than the others, and the median 30% lower than the FCFS heuristic, lowest in the literature. These results comes from the fact that evaluating the whole set of jobs, GA-BL has the ability to forbid access to some nodes by means of the blacklist, providing a reservation mechanism for those jobs with computational requirements that can provide better benefits in the future.

The results of the experimentation for the RICC workload can be seen in Figure 4. The behavior of GA-BL was similar to the previous experiment. It gave the best results but

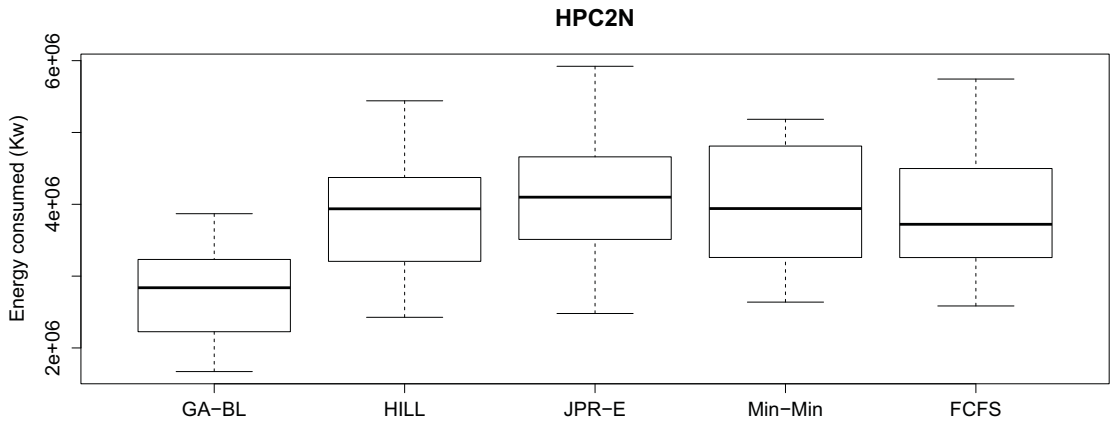


Figure 3: Energy saving study for HPC2N workload

also the lowest deviation of the boxplot, meaning that it obtains the lowest consumption irrespectively of the sets of jobs. We can also observe that the outliers corresponding to the sets with higher consumption than the median have also have much better results in our proposal than in the other techniques.

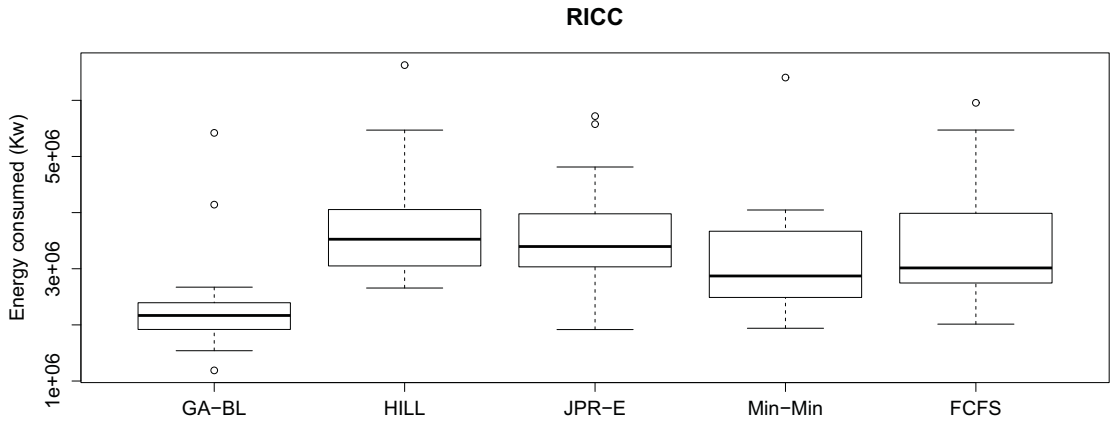


Figure 4: Energy saving study for RICC workload

In the case of the CEA CURIE, the workloads results are shown in Figure 5. The median and boxplot results are very similar for all the techniques. The Min-Min is the only that differs slightly from the others, being the worst technique.

It is important to point out that the jobs in this workload have very low execution times. This means that the differences in scheduling decisions about resources allocation or job ordering have little effect on the consumption results.

Nonetheless, it should be noted that the GA-BL presents a higher rate of sets in the first quartile, which implies more energy savings.

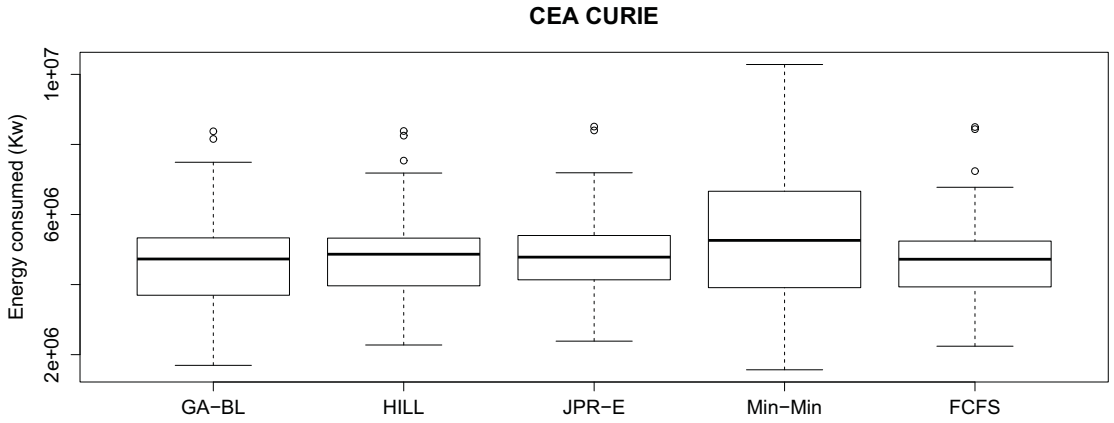


Figure 5: Energy saving study for CEA CURIE workload

We can conclude that the workloads with higher variability in the execution times (HPC2N and RICC) produce better scheduling chances. Our proposal, GA-BL technique, using a blacklist of computational resources has proved to be effective for the energy saving and also shows a low sensitivity to workloads variations. However, for the situations where the nature of the workloads does not provide scheduling chances, as in the case of the CEA CURIE, all the techniques gave similar results. Nevertheless, our proposal managed to obtain higher rate of scheduled workloads with lower energy consumption.

5 Conclusions

The paper proposes a novel scheduling technique based on a resource allocation blacklist. This proposal avoids the systematically allocation of the jobs to the resources with minimum consumption, providing new scheduling opportunities to reduce global consumption. Our proposal is integrated into a Genetic Algorithm based meta-heuristic (GA-BL) able to modify the job execution order based on the blacklist.

The performance of the GA-BL in energy saving was evaluated by simulation and compared with other techniques in the literature. The experimental study was conducted with real workload traces in a heterogeneous federated cluster environment.

The empirical results show that GA-BL can significantly reduce energy consumption and have also shown a low sensibility to workloads variations. Although in situations where the results were similar among all the techniques because of the limited opportunities to reduce the energy, the GA-BL obtained higher rate of scheduling results with the lowest energy consumption.

Acknowledgements

This research is partly supported by the European Union FEDER (CAPAP-H5 network TIN2014-53522-REDT) and MEyC-Spain under contract TIN2014-53234-C2-2-R

References

[1] Carsten Ernemann, Volker Hamscher, Uwe Schwiegelshohn, Ramin Yahyapour, and Achim Streit. On advantages of grid computing for parallel job scheduling. In *CCGRID,IEEE*, pages 39–39, 2002.

[2] Anca I.D. Bucur and Dick H.J. Epema. Scheduling policies for processor coallocation in multicluster systems. *Transactions on Parallel and Distributed Systems, IEEE*, 18(7):958–972, 2007.

[3] Hector Blanco, Josep L. Lerida, Fernando Cores, and Fernando Guirado. Multiple job co-allocation strategy for heterogeneous multi-cluster systems based on linear programming. *The Journal of Supercomputing*, 58(3):394–402, 2011.

[4] Dongbo Liu and Ning Han. Co-scheduling deadline-sensitive applications in large-scale grid systems. *International Journal of Future Generation Communication & Networking*, 7(3):49–60, 2014.

[5] Shekhar Srikantaiah, Aman Kansal, and Feng Zhao. Energy aware consolidation for cloud computing. In *Proceedings of the 2008 conference on Power aware computing and systems*, volume 10, pages 1–5. San Diego, California, 2008.

[6] Chia-Hung Lien, Ming Fong Liu, Ying-Wen Bai, Chi Hsiung Lin, and Ming-Bo Lin. Measurement by the software design for the power consumption of streaming media servers. In *Instrumentation and Measurement Technology Conference, 2006. IMTC 2006. Proceedings of the IEEE*, pages 1597–1602. IEEE, 2006.

[7] Alberto Cocaña, José Ranilla, and Luciano Sánchez. Energy-efficient allocation of computing node slots in HPC clusters through parameter learning and hybrid genetic fuzzy system modeling. *The Journal of Supercomputing*, 71(3):1163–1174, 2015.

[8] Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, and Ronald P. Doyle. Managing energy and server resources in hosting centers. *SIGOPS Oper. Syst. Rev.*, 35(5):103–116, October 2001.

[9] Eduardo Pinheiro, Ricardo Bianchini, Enrique V. Carrera, and Taliver Heath. Load balancing and unbalancing for power and performance in cluster-based systems. In *Workshop on compilers and operating systems for low power*, volume 180, pages 182–195. Barcelona, Spain, 2001.

[10] Anne-Ccile Orgerie, Laurent Lefevre, and Jean-Patrick Gelas. Save watts in your grid: Green strategies for energy-aware framework in large scale distributed systems. In *Parallel and Distributed Systems, 2008. ICPADS '08. 14th IEEE International Conference on*, pages 171–178, Dec 2008.

[11] Yu Li, Yi Liu, and Depei Qian. A heuristic energy-aware scheduling algorithm for heterogeneous clusters. In *Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on*, pages 407–413, Dec 2009.

[12] Sandy Irani, Sandeep Shukla, and Rajesh Gupta. Online strategies for dynamic power management in systems with multiple power-saving states. *ACM Transactions on Embedded Computing Systems (TECS)*, 2(3):325–346, 2003.

[13] Ching-Hsien Hsu, Kenn D. Slagter, Shih-Chang Chen, and Yeh-Ching Chung. Optimizing energy consumption with task consolidation in clouds. *Information Sciences*, 258:452 – 462, 2014.

[14] Eloi Gabaldon, Josep L. Lerida, Fernando. Guirado, and Jordi Planes. Multi-criteria genetic algorithm applied to scheduling. *J Simulation*, pages –, 2015.

[15] Vijay K. Naik, Chuang Liu, Lingyun Yang, and Jonathan Wagner. Online resource matching for heterogeneous grid environments. In *CCGRID*, pages 607–614, 2005.

[16] Dror Feitelson. Parallel workloads archive. http://www.cs.huji.ac.il/labs/parallel/workload, 2005.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Parallel processing in GPUs for intra-picture prediction in HEVC

Vicente Galiano[1], Héctor Migallón[1], Victoria Herranz[2], Pablo Piñol[1], Otoniel López-Granado[1] and Manuel P. Malumbres[1]

[1] *Department of Physics and Computer Architecture, Miguel Hernández University*

[2] *Center of Operations Research, Miguel Hernández University*

emails: vgaliano@umh.es, hmigallon@umh.es, mavi.herranz@umh.es, pablop@umh.es, otoniel@umh.es, mels@umh.es

Resumen

The HEVC video coding standard launched on 2013, is able to reduce to the half, on average, the bit stream size produced by H.264/AVC encoder at the same video quality, but it requires nearly 70 % more time than H.264/AVC to encode a video sequence. GPUs can help to reduce this coding time considerably. In this paper, we propose the use of GPUs to perform the intra-picture prediction, explaining which steps in the coding process has been traslated to GPU and we compare the coding time with respect to the intra-picture prediction computation on a CPU.

Key words: Parallel algorithms, video coding, HEVC, GPUs, performance, prediction

1. Introduction

HEVC, the High Efficiency Video Coding standard [1] launched on January 2013 by the Joint Collaborative Team on Video Coding (JCT-VC) is the newest video coding standard of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). The main video coding standard preceding HEVC is the current H.264/AVC [2] standard, but an increasing diversity of services and the emergence of 4K and 8K resolution are creating stronger needs for better coding efficiency. HEVC greatly improved this coding efficiency over its predecessor (H.264/AVC) by a factor of almost twice while maintaining an equivalent visual quality [3]. HEVC has been designed to cover all services of H.264/AVC and to focus on two key issues: increased video resolution and

increased use of parallel processing architectures. In terms of complexity, Bossen et al. [4] studied the complexity aspects of HEVC encoding and decoding software and concluded that the encoding process is much more challenging than the decoding process.

Despite being a recent standard, we can find in the literature several works about complexity analysis and parallelization strategies for the HEVC standard [4, 5, 6]. Most of the parallelization proposals are focused in the decoding side, looking for the most appropriate parallel optimizations at the decoder that provide real-time decoding of High-Definition (HD) and Ultra-High-Definition (UHD) video contents. In [7] and [8] the authors present a technique called Overlapped Wavefront (OWF) for the HEVC decoder which is a variant of Wavefront Parallel Processing (WPP) in which the executions over consecutive pictures are overlapped. In a multi-threaded approach of the HEVC decoder, a picture is decoded by several threads at the same time, and each thread decodes different Coding Tree Block (CTB) rows. In these works, authors claim that a single thread may continue processing the next picture when it finishes the current one, without waiting for the other threads. These variations allow a better parallel processing efficiency, reducing the overall decoding time. Recently, in [9] the authors combine Tiles, WPP with SIMD (Single-Instruction Multiple-Data instruction set extension to the x86 architecture) instructions to develop a real-time HEVC decoder.

However, there are less works focused at the HEVC encoder. In [10] authors propose a fine-grain parallel optimization in the motion estimation module of the HEVC encoder allowing to perform the motion vector prediction in all Prediction Units (PUs) available at the Coding Unit (CU) at the same time.

In [11], authors apply parallel processing techniques to HEVC encoder. Authors propose several, synchronous and asynchronous, parallelization approaches working at a coarse grain parallelization level, based on the Group Of Pictures (GOP), where several groups of consecutive frames are encoded simultaneously using a multicore platform with shared memory. The results show that near ideal efficiencies are obtained using up to 10 cores.

A proposal for parallelization for distributed memory systems is presented in [12]. In this paper, authors present several parallelization approaches to the HEVC encoder for distributed memory platforms which work at a coarse grain level parallelization, being one group of pictures (GOP) the basic structure. These approaches encode simultaneously several GOPs and ideal parallel behavior is shown for a right GOP conformation and distribution.

In [13], authors combine a GPU-based motion estimation algorithm with two different parallelization techniques: WPP and group of pictures (GOP). This approach allows a multicore system to process multiple Coding Tree Units (CTUs) by splitting the frame in rows or the sequence in GOPs, respectively. In either case, the motion estimation of these regions is issued to the GPU device obtaining speed-ups of up to 3.93x for 4 processes.

In [14] authors propose a parallelization inside the intra prediction module that consist on removing data dependencies among subblocks of a CU, obtaining interesting speed-up

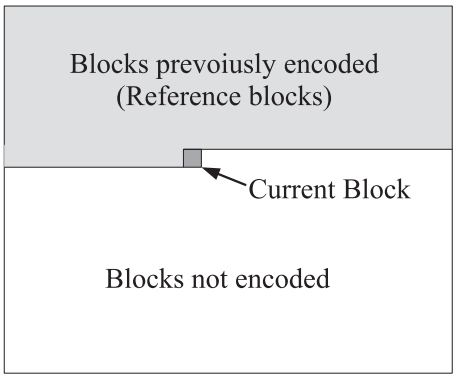


Figura 1: Avalaible blocks on decoder side for intra-prediction

results.

Recently, in [15], authors present an hybrid approach combining WPP and intra prediction on GPUs obtaining reductions on the total encoding time of up to 62 % with a lower coding performance loss.

In this paper, we will focus on applying parallel processing in Graphic Processors Units (GPUs) to the intra-picture prediction process of the HEVC encoder. The remainder of this paper is organized as follows, in Section 2 an overview of intra-picture prediction in HEVC and an introduction to the main algorithms used are presented. Section 3 present the parallelization strategy proposed for using GPUs parallelism in the intra-picture prediction process, while in Section 4 an evaluation of the proposed architecture and parallel strategies is presented. Finally, in Section 5 some conclusions are drawn.

2. Intra-picture Prediction in HEVC

The basic source-coding algorithm is a hybrid of interpicture prediction to exploit temporal statistical dependences, intra-picture prediction to exploit spatial statistical dependences, and transform coding of the prediction residual signals to further exploit spatial statistical dependences. These three elements provide an improvement in compression efficiency when compared to the previous video coding standard H.264/AVC. In particular, the intra-picture prediction process consists in predicting a block in the current frame, using the information from neighbouring blocks in the same frame. It supports three different modes, the angular mode with 33 different directions, the planar mode and the DC mode.

The values of the samples in a frame are often similar to their adjacent neighbour samples' values; this is called spatial redundancy or intra-frame correlation. This redundant information in the spatial domain can be exploited to compress the image. Each frame is partitioned in blocks and for each block the prediction is created by extrapolating samples

from previously-coded samples in the same frame. In order for this process to be replicable at the decoder side, only the pixels along the upper and/or left edges can be used to create the prediction block as shown in Figure 1. Once the prediction has been generated, it is subtracted from the current block to form a residual signal. The residual signal is transformed into the frequency domain and binary arithmetic encoded, together with the selected prediction mode.

In HEVC, a frame is split into one or several slices and an intra slice contains a number of consecutive CTUs, which are partitioned into CUs. The maximum CU size is 64x64, and the minimum size is 8x8. A CU is considered as whole or partitioned into 4 smaller CUs. Whether to further split the current CU depends on its Rate-Distortion (RD) cost and the total RD cost of the 4 smaller CUs.

HEVC employs 35 different intra modes to predict a CU, compared to the 8 modes available in H.264/AVC. The video encoder will choose the intra prediction mode that provides de best RD performance. The prediction modes are organized into three categories:

- Planar prediction: the value of each sample of the prediction CU is calculated assuming an amplitude surface with a horizontal and vertical slope derived from the boundary samples of the neighbouring blocks (mode 0) .
- DC prediction: the value of each sample of the prediction CU is an average of the boundary samples of the neighbouring blocks (mode 1).
- Directional prediction with 33 different directional orientations: the value of each sample of the prediction CU is calculated extrapolating the value from the boundary samples of the neighbouring blocks as shown in Figure 2 (mode $2 \dots 34$).

The residual block is obtained as the difference between the original CU and the prediction CU, or:

$$residualAngularBlock = OriginalCU - PredictionAngCU$$

Finally the Sum of Absolute Differences (SAD) of the residual block is calculated as:

$$SADmodeAngular = sum(abs(residualAngularBlock))$$

In HEVC, the increase in the number of intra prediction modes provides substantial coding performance gain over H.264/AVC, but it also makes the RD optimization process more complex. The fast encoding algorithm of HEVC reference software includes two phases.

In the first phase, called Rough Mode Decision (RMD), the N most promising candidate modes are selected. In this process, all candidates (35 modes) are evaluated with respect to the following Hadamard transform difference (Had) cost function:

$$C = D_{Had} + \lambda \cdot R_{mode}$$

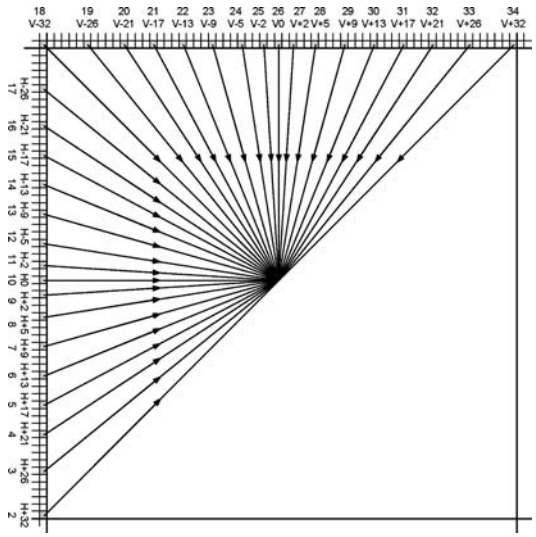


Figura 2: HEVC angular intra prediction modes

where the D_{Had} represents the absolute sum of Hadamard transformed residual signal for a CU, λ is the Lagrange multiplier that determines the trade-off between rate and distortion, and R_{mode} represents the number of bits of the block when it is encoded with CABAC. Then, up to three modes with the lowest costs are added to a subset of candidates (SC).

In the second phase, the full RD optimization process is performed on all the candidates in set SC, and the intra prediction mode with the minimum RD cost is selected. The total complexity of this step depends on the number of modes in set SC.

3. Moving Hadamard transformed to GPUs

The release of NVIDIA CUDA API [16, 17, 18] to developers has led to an spectacularly increase of interest in using the GPU capabilities towards faster and more efficient computation in parallel. The GPU serves as a coprocessor to the CPU through the CUDA API and exploits the massive data parallelism on the Single Instruction Multiple Data (SIMD) architecture of the GPU. Independently operating threads executing CUDA kernels while efficiently sharing high speed memory can be implemented with a set of threads being organized into blocks. It should be noted that to obtain higher bandwidth and overall performance gains, memory sharing between threads must be optimized with very careful programming to ensure the least amount of memory latency between reads/writes.

At the moment, we have introduced the algorithm used in the intra-picture prediction which is based in a Single Instruction Single Data (SISD) architecture. SAD computing is

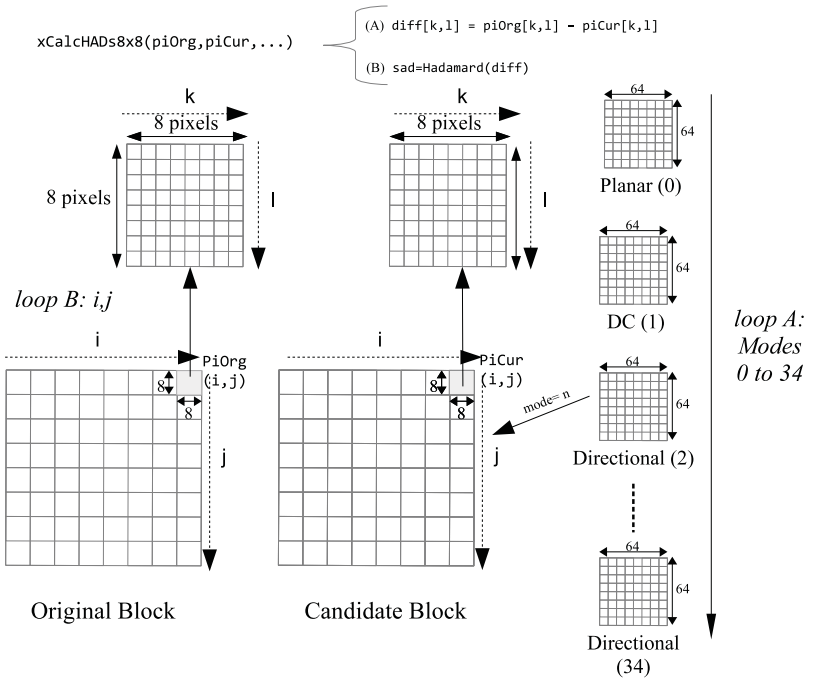


Figura 3: Algorithm for calculate D_{Had} for each mode

implemented in a CB of size 8×8 pixels so for each CB, differents SAD values must be calculated 8×8 times in a CTU of size 64×64, and $\times 35$ times corresponding to the 35 prediction modes. In this section, we explain how Hadamard transformed is implemented in the HM 16.3 reference software, and how we have moved this massive computing to GPUs in order to accelerate the prediction algorithm using the data parallelism.

In Figure 3, a representation of the algorithm used for getting the SAD for each mode is illustrated. First, we have the original block on the left and the best candidate block must be searched for modes 0 to 34 (loop A). For the candidate block from mode n, the difference from original and the Hadamard transformed are obtained in blocks of 8×8 pixels, so we must get the transformed to 64 subblocks of size 8×8 (loop B). In the HM 16.3 reference software, the function `xCalcHAD8x8` is called for each original subblock (i, j) and each candidate subblock (i, j) with size 8×8. This function has two steps: first, the difference between predicition and original pixels is calculated; second, the Hadamard transformed from the difference matrix of size 8×8 is obtained. Inside this function, a new iteration (loop C) is implemented to calculate the difference for each element (k, l) and the Hadamard transformed of the difference matrix. As we can note, there are not data dependency between adjacent subblocks and all SAD values for each CTU could be

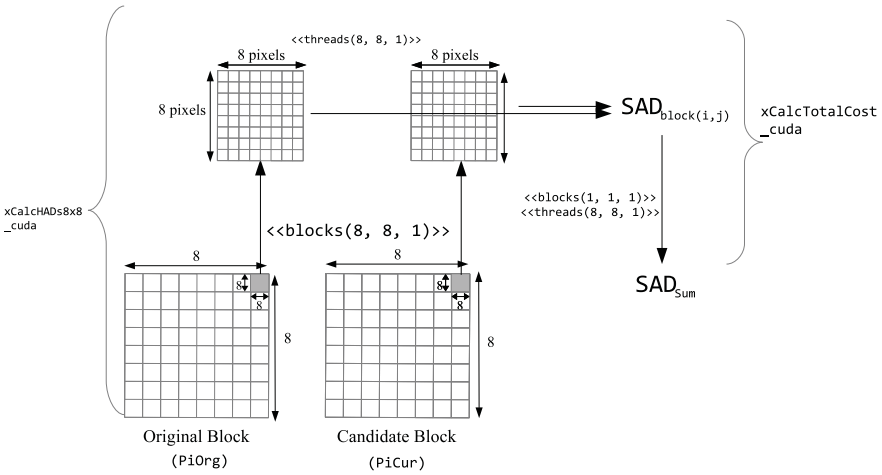


Figura 4: Concurrent SAD computation in GPU

calculated simultaneously. But in each subblock there are dependencies on the Hadamard transformed computing, and there are dependencies in the necessary reduction processes to compute the sublock differences.

In this work we propose the concurrent calculation of SAD values (including the Hadamard transformed) for all subblocks belonging to the CTU. In Figure 4, the concurrent algorithm to calculate the SAD in GPU is showed. The sum of SAD values is obtained in two steps. In a first step, we define a kernel grid of 8×8 blocks of 8×8 threads. Each block of threads must calculate a SAD value for each subblock. Each thread participate in the calculation of its difference value and Hadamard transformed is obtained in a distributed computing between the threads of the block. At the end of this first step, we get a vector of 64 elements with the SAD value for each subblock. Due to each value has been computed by a block of threads, to sum these values, a new kernel in the GPU is called performing an optimized reduction process based on the use of shared memory. The result is the SAD_{Sum} associated with the mode n. We must note that before the first computation step, both CTUs (original and candidate) of size 64×64 must be transfered from host memory to the global memory of the device using asyncrhonous transfers. Transfer times can be overlapped with computing times for previous CTUs. Thereby, communication times between global memory and device memory should not represent a significant penalty. On the other hand, the computation of sum values in the first and second steps is implemented using the shared memory. This shared memory is allocated per thread block, so all threads in the block have access to the same shared memory, which latency is roughly $100x$ lower than uncached global memory latency.

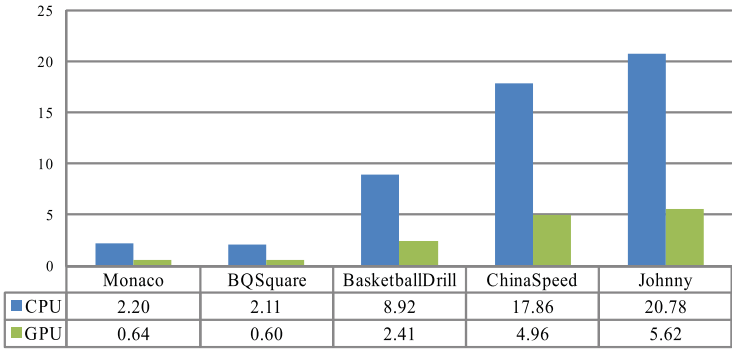
4. Performance Evaluation

In this section we present the performance evaluation of our GPU-based SAD computing algorithm in terms of computational times obtained and we compare them with the ones obtained in a CPU using the sequential algorithm of the reference software. Two different platforms have been used in this work. The first one is a Nvidia Tesla M2050 which contains 448 CUDA cores with 3 GB of dedicated video memory. The second one is a laptop GPU Geforce GT540M with 96 CUDA cores and 2 GB of video memory. The sequential algorithm has been also executed in two platforms: first, a node with two processors Intel Xeon X5660 and 48 GB of RAM memory and second a labtop with an Intel i7-2670QM at 2.2GHz and 8GB of RAM memory.

In Figure 5, we present the computational times required for computing SADs in both GPU platforms used in this work (Tesla M2050 and Gforce GT540M, respectively) and for five different video frame resolutions: Monaco (352×288), BQSquare (416×240) , BasketballDrill (832×480), ChinaSpeed (1024×768) and Johnny (1280×720). In all video sequences, we have encoded 50 frames. In both figures, *CPU* means the sequential computational time required to compute the SAD values and *GPU* indicates the computational time for the proposed GPU based algorithm. As we can observe, we obtain remarkables time reductions, above 70 % for the M2050 and above 60 % for the GT540M. Note that the timer reduction does not depend on the resolution image. Obviously, due to M2050 is more powerful than the GT540M, we obtain better computational results using the first one. On the other hand, note tha both the Hadamard transformed and the SAD compunting include reduction operations. These reduction operations decrease substantially the inherent parallelism. To be exploited, we execute these operations on the GPUs managing efficiently the shared memory abd mapping suitably the kernel grid.

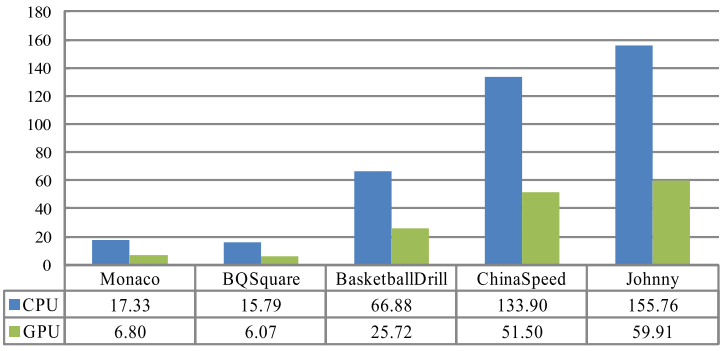
5. Conclusions and Future Work

In this paper we have proposed the use of GPUs for computing the SAD values used in intra-picture prediction in HEVC. The search of the best prediction block implies a gready search of candidates among 35 modes for each block. On the other hand, this prediction algorithm is massively used in all CTUs that belongs to a frame. Besides, we have detailed how the algorithm is implemented in the 16.3 HM reference software and how concurrency can be highly exploited when computing the sum of CTU's SAD, spliting it in blocks and threads. Considering that the time reduction are up to 73 %, we value these resultas as a good first step to substantially improve the intra-picture prediction algorithm using GPUs. For future work, we plan to increase the concurrent task in GPUs computing all modes at the same time. Even further work will lead us to make a prediction of the entire frame concurrently.



	Monaco	BQSquare	BasketballDrill	ChinaSpeed	Johnny
CPU	2.20	2.11	8.92	17.86	20.78
GPU	0.64	0.60	2.41	4.96	5.62

(a) Nvidia Tesla M2050



	Monaco	BQSquare	BasketballDrill	ChinaSpeed	Johnny
CPU	17.33	15.79	66.88	133.90	155.76
GPU	6.80	6.07	25.72	51.50	59.91

(b) Nvidia GForce GT540M

Figura 5: Computational times for GPU-based Hadamard transformed for intra-picture prediction in HEVC

Acknowledgments

This research was supported by the Spanish Ministry of Economy and Competitiveness under Grant TIN2015-66972-C5-4-R co-financed by FEDER funds.

Referencias

[1] B. Bross, W. Han, J. Ohm, G. Sullivan, Y.-K. Wang, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 10," *Document JCTVC-L1003 of JCT-VC, Geneva*, January 2013.

[2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," *ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC) version 16, 2012*, 2012.

[3] G. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *Circuits and systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1648 –1667, December 2012.

[4] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC complexity and implementation analysis," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1685–1696, 2012.

[5] M. Alvarez-Mesa, C. Chi, B. Juurlink, V. George, and T. Schierl, "Parallel video decoding in the emerging HEVC standard," in *International Conference on Acoustics, Speech, and Signal Processing, Kyoto*, March 2012, pp. 1–17.

[6] E. Ayele and S.B.Dhok, "Review of proposed high efficiency video coding (HEVC) standard," *International Journal of Computer Applications*, vol. 59, no. 15, pp. 1–9, 2012.

[7] C. C. Chi, M. Alvarez-Mesa, B. Juurlink, G. Clare, F. Henry, S. Pateux, , and T. Schierl, "Parallel scalability and efficiency of HEVC parallelization approaches," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1827 –1838, 2012.

[8] C. C. Chi, M. Alvarez-Mesa, J. Lucas, B. Juurlink, and T. Schierl, "Parallel HEVC decoding on multi- and many-core architectures," *Journal of Signal Processing Systems*, vol. 71, no. 3, pp. 247 –260, 2013.

[9] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, Y.-K. Wang, and T. Wiegand, "High Efficiency Video Coding (HEVC) text specification draft 10," Joint Collaborative Team on Video Coding (JCT-VC), Geneva (Switzerland), Tech. Rep. JCTVC-L1003, January 2013.

[10] Q. Yu, L. Zhao, and S. Ma, "Parallel AMVP candidate list construction for HEVC," in *VCIP'12*, 2012, pp. 1–6.

[11] H. Migallón, J. Hernández-Losada, G. Cebrián-Márquez, P. Piñol, J. Martínez, O. López-Granado, and M. Malumbres, "Synchronous and asynchronous {HEVC} parallel encoder versions based on a {GOP} approach," *Advances in Engineering Software*, pp.–, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S096599781630028X

[12] H. Migallón, V. Galiano, P. Piñol, O. López-Granado, and M. P. Malumbres, "Distributed memory parallel approaches for hevc encoder," *The Journal of Supercomputing*, pp. 1–12, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11227-016-1666-2

[13] G. Cebrián-Márquez, J. L. Hernández-Losada, J. L. Martínez, P. Cuenca, M. Tang, and J. Wen, "Accelerating HEVC using heterogeneous platforms," *Journal of Supercomputing*, vol. 71, no. 2, pp. 613 –628, February 2015.

[14] J. Jiang, B. Guo, W. Mo, and K. Fan, "Block-based parallel intra prediction scheme for HEVC," *Journal of Multimedia*, vol. 7, no. 4, pp. 289 –294, August 2012.

[15] S. Radicke, J. U. Hahn, Q. Wang, and C. Grecos, "A parallel hevc intra prediction algorithm for heterogeneous cpu+gpu platforms," *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 103–119, March 2016.

[16] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with cuda," in *Queue*, vol. 6, no. 2, 2008, pp. 40–53.

[17] N. Corporation, "Nvidia cuda c programming guide. version 3.2."

[18] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "Nvidia tesla: A unified graphics and computing architecture," in *IEEE Micro*, vol. 28, no. 2, 2008, pp. 39–55.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A modified exponential method to approximate positive and bounded solutions of the Burgers-Fisher equation

Armando Gallegos[1], J. E. Macías-Díaz[2] and H. Vargas-Rodríguez[1]

[1] *Centro Universitario de los Lagos, Universidad de Guadalajara*

[2] *Departamento de Matemáticas y Física, Universidad Autónoma de Aguascalientes*

emails: gallegos@culagos.udg.mx, jemacias@correo.uaa.mx,
hvargas@culagos.udg.mx

Abstract

In this work, we propose an exponential-type discretization of the well-known Fisher's equation from population dynamics. Only positive and bounded solutions are physically relevant in this note, and the discretization that we provide is able to preserve both properties. The method is a novel explicit exponential technique which has the advantage of requiring less computational resources and less computer time. It is worthwhile to note that our technique has the advantage over other exponential methodologies that it yields no singularities. In addition, the preservation of the properties of non-negativity and boundedness are distinctive features of our method.
Key words: Modified Bhattacharya method, non-negativity, boundedness, efficiency

1 Introduction

Throughout, we let α and β be positive numbers, and let Ω be a domain of $\mathbb{R}^2$ which is open, bounded and connected. We let $u = u(x, y, t)$ be a real function defined on the closure of $\Omega \times \mathbb{R}^+$, which is twice differentiable in the interior of its domain and which satisfies the initial-boundary-value problem

$$\begin{aligned} \frac{\partial u}{\partial t}(x,y,t) &= \Delta u(x,y,t) + \alpha u(x,y,t)\left(1-\beta u(x,y,t)\right), \quad \forall (x,y)\in\Omega, \forall t\in\mathbb{R}^+, \\ &\begin{cases} u(x,y,0) = \phi(x,y), & \forall (x,y)\in\overline{\Omega}, \\ \dfrac{\partial u}{\partial \mathbf{n}}(x,y,t) = 0, & \forall (x,y)\in\partial\Omega, \forall t\in\mathbb{R}^+\cup\{0\}, \end{cases} \end{aligned} \tag{1}$$

 ISBN: 978-84-608-6082-2

for some continuous function $\phi : \overline{\Omega} \to \mathbb{R}$ that satisfies $0 \leq \phi(x,y) \leq \frac{1}{\beta}$ at each $(x,y) \in \overline{\Omega}$. Here, Δ represents the Laplacian operator on the spatial variables x and y. Clearly, the partial differential equation of (1) is the two-dimensional generalization of the diffusion-reaction equation investigated simultaneously and independently in 1937 by R. A. Fisher [1] and A. N. Kolmogorov, I. G. Petrovsky and N. S. Piscounov [2].

Let κ be a positive number. It is well-known that the two-dimensional Fisher's equation has non-negative and bounded solutions, some of them traveling waves [3]. In view of this fact, we will restrict our attention to solutions satisfying $u(x,y,t) \geq 0$ for each $(x,y) \in \overline{\Omega}$ and $t \geq 0$. After dividing both sides of (1) by the nonnegative function $u(x,t,y) + \kappa$ and using the chain rule, we obtain

$$\frac{\partial \ln(u(x,y,t)+\kappa)}{\partial t} = \frac{\Delta u(x,y,t) + \alpha u(x,y,t)\,(1-\beta u(x,y,t))}{u(x,y,t)+\kappa}, \tag{2}$$

for each $(x,y) \in \Omega, \forall t \in \mathbb{R}^+$. Associated to this differential equation, we consider the same initial-boundary conditions as in the problem (1) for a continuous and nonnegative function ϕ. In the present work, we are interested in developing a numerical method to approximate the solution of (1), with the following characteristics: **(a)** the non-negativity and the boundedness of approximations are preserved, **(b)** the technique is computationally fast, **(c)** the method is easy to implement in any computer language, and **(d)** the computational implementation allows to employ fine grid meshes.

2 Exponential method

Let M and N be positive integers and suppose that Ω is the rectangle $[a,b] \times [c,d]$ of $\mathbb{R}^2$, where $a < b$ and $c < d$. Fix uniform partitions $\{x_m\}_{m=0}^{M}$ and $\{y_n\}_{n=0}^{N}$ of $[a,b]$ and $[c,d]$, respectively, with step-sizes given by Δx and Δy. Let $\{t_k\}_{k=0}^{\infty}$ be a partition of the temporal interval $[0,\infty)$. For each $k \in \mathbb{Z}^+ \cup \{0\}$ and $z = x, y$, let

$$\Delta t_k = t_{k+1} - t_k, \tag{3}$$

$$R_z = \frac{\Delta t_k}{(\Delta z)^2}. \tag{4}$$

Let $w_{m,n}^k$ represent an estimate of $u(x_m, y_n, t_k)$, for each $m \in \{0,1,\ldots,M\}$, $n \in \{0,1,\ldots,N\}$ and $k \in \mathbb{Z}^+ \cup \{0\}$. Throughout, we use the following operators:

$$\delta_t w_{m,n}^k = \frac{w_{m,n}^{k+1} - w_{m,n}^k}{\Delta t_k}, \tag{5}$$

$$\delta_{xx} w_{m,n}^k = \frac{w_{m+1,n}^k - 2w_{m,n}^k + w_{m-1,n}^k}{(\Delta x)^2}, \tag{6}$$

$$\delta_{yy} w_{m,n}^k = \frac{w_{m,n+1}^k - 2w_{m,n}^k + w_{m,n-1}^k}{(\Delta y)^2}, \tag{7}$$

for each $m \in \{1,\ldots,M-1\}$, $n \in \{1,\ldots,N-1\}$ and $k \in \mathbb{Z}^+ \cup \{0\}$. Obviously, these operators approximate the values of the functions u_t, u_{xx} and u_{yy} at the point (x_m, y_n, t_k) with order of consistency equal to Δt, $(\Delta x)^2$ and $(\Delta y)^2$, respectively. Finally, we will impose exact discrete conditions at the time $t = 0$, and discrete homogeneous Neumann conditions at the boundary of the spatial domain.

Let $m \in \{1,\ldots,M-1\}$, $n \in \{1,\ldots,N-1\}$ and $k \in \mathbb{Z}^+ \cup \{0\}$, and let $(\kappa_k)_{k=0}^{\infty}$ be a sequence of positive numbers. We discretize the partial differential equation (2) at the point (x_m, y_n, t_k) as follows:

$$\frac{\ln(w_{m,n}^{k+1}+\kappa_k) - \ln(w_{m,n}^{k}+\kappa_k)}{\Delta t_k} = \frac{(\delta_{xx}+\delta_{yy})\, w_{m,n}^{k} + \alpha w_{m,n}^{k}(1-\beta w_{m,n}^{k})}{w_{m,n}^{k}+\kappa_k}. \tag{8}$$

Equivalently,

$$w_{m,n}^{k+1} = (w_{m,n}^{k}+\kappa_k)\exp\left[\frac{\Delta t_k\,(\delta_{xx}+\delta_{yy})\, w_{m,n}^{k} + \alpha\Delta t_k w_{m,n}^{k}(1-\beta w_{m,n}^{k})}{w_{m,n}^{k}+\kappa_k}\right] - \kappa_k, \tag{9}$$

which evinces the explicit nature of (8). An alternative expression of this method is readily at hand if we consider the following notation. Let

$$A_k = \alpha\beta\Delta t_k, \tag{10}$$

$$B_k = \alpha\Delta t_k + 2(R_x^k + R_y^k), \tag{11}$$

$$C_{m,n}^k = R_x^k(w_{m+1,n}^k + w_{m-1,n}^k) + R_y^k(w_{m,n+1}^k + w_{m,n-1}^k). \tag{12}$$

Clearly, the method (8) can be rewritten as $w_{m,n}^{k+1} = F_{m,n}^k(w_{m,n}^k)$, where

$$F_{m,n}^k(w) = (w+\kappa_k)\exp\left[\frac{A_k w^2 - B_k w + C_{m,n}^k}{w+\kappa_k}\right] - \kappa_k. \tag{13}$$

Before closing this section we would like to emphasize the simplicity of the computational implementation of this scheme. There are various reports in the literature which describe discretizations similar to (8), most notably [4, 5]. However, those approaches use values of $\kappa_k = 0$ for each $k \in \mathbb{Z}^+ \cup \{0\}$. Those techniques do not allow for numerical solutions to be close or equal to zero at any point of the grid. From that perspective, the inclusion of the parameter κ in the finite-difference discretization (8) avoids divisions by zero when the approximations $w_{m,n}^k$ are allowed to take on that value.

3 Dynamical properties

We establish here conditions that guarantee the positivity and the boundedness of approximations obtained through (8). To that end, it suffices to bound the range of the function

 ISBN: 978-84-608-6082-2

$F^k_{m,n} : [0, \frac{1}{\beta}] \to \mathbb{R}$ of (13) in $[0, \frac{1}{\beta}]$. We will restrict our attention to cases when $\Delta x = \Delta y = 1$. Also, we will use $\mathbf{w}^k$ to represent the lexicograpically ordered vector of approximations at the time t_k, for $k \in \mathbb{Z}^+ \cup \{0\}$.

Lemma 1 *Let $k \in \mathbb{Z}^+ \cup \{0\}$, and suppose $\Delta x = \Delta y = 1$ and $\Delta t_k(\alpha + 4) < 1$. Then the function $F^k_{m,n}$ of (13) is increasing in $[0, \frac{1}{\beta}]$ when $0 \leq \mathbf{w}^k \leq \frac{1}{\beta}$ and*

$$\kappa_k > \frac{4\Delta t_k}{\beta[1 - \Delta t_k(\alpha + 4)]}. \tag{14}$$

The following is the main result on the existence and uniqueness of positive and bounded solutions of (8).

Proposition 2 *Let $0 \leq \mathbf{w}^0 \leq \frac{1}{\beta}$, $\Delta x = \Delta y = 1$, $\Delta t_k(\alpha + 4) < 1$, and suppose that (14) holds for each $k \in \mathbb{Z}^+ \cup \{0\}$. Then there exists a unique sequence $(\mathbf{w}_k)_{k=0}^{\infty}$ satisfying the method (8) and the discrete Neumann boundary conditions, such that $0 \leq \mathbf{w}^k \leq \frac{1}{\beta}$ holds for each $k \in \mathbb{Z}^+ \cup \{0\}$.*

In the present work, these and more analytical/numerical results will be extended to the case of the Burgers-Fisher equation from population dynamics.

References

[1] Ronald Aylmer Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937.

[2] Andrei N Kolmogorov, IG Petrovsky, and NS Piskunov. Etude de léquation de la diffusion avec croissance de la quantité de matiere et son applicationa un probleme biologique. *Mosc. Univ. Bull. Math*, 1:1–25, 1937.

[3] Mark J Ablowitz and Anthony Zeppetella. Explicit solutions of Fisher's equation for a special wave speed. *Bulletin of Mathematical Biology*, 41(6):835–840, 1979.

[4] A Refik Bahadır. Exponential finite-difference method applied to Korteweg–de Vries equation for small times. *Applied Mathematics and Computation*, 160(3):675–682, 2005.

[5] Bilge Inan and Ahmet Refik Bahadir. A numerical solution of the Burgers equation using a Crank-Nicolson exponential finite difference method. *Journal of Mathematical and Computational Science*, 4(5):849–860, 2014.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Invariant set for third-order switched systems

J. B. García-Gutiérrez[1], F. Benítez-Trujillo[1] and C. Pérez[1]

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

emails: juanbosco.garciagutierrez@alum.uca.es, quico.benitez@uca.es, carmen.perez@uca.es

Abstract

We study invariant sets for switched systems. We focus on a class of third-order switched systems. In this class of switched systems we solve the problem of finding an invariant set for a switching law. For each switched system we provide the invariant set and the swithching law, in addition, the invariant set is a polyhedral cone and the switching law is switching on the boundary. To accomplish this we reduce the problem to a simplified system. We provide a procedure calculating the invariant set. The method is based in calculaing an invariant set for a simplified system and then transformating the invariant set to the original system. We illustrate this method with an example.

Key words: Invariant set, switched linear systems, third-order system

1 Introduction

A switched linear system is a dynamical system which consist of several subsystem, and one of them is active each time. There are two posibilities of studying properties of a switched system, one way is when a property holds for every switching law and other way is when a property holds for at least one switching law and in this last case the problem is to design a switching law. For instance, stability is studied in this two ways, see the survey paper [1].

In this paper we only consider continuous switched system, although there are examples and papers on discrete switched systems [2]. A wide number of properties for dynamical systems have been studied for switched system as well. For example controllability and reachability [3], stability and stabilization [1], optimization [4]. See the book [5] to find others problems for switched systems.

We study the problem of invariant set for switched systems. Invariant sets for dynamical system have been studied for example in [6]. Invariant set in the scope of switched system

has been given in the paper [7] but in the context of arbitrary switching law. However we study when a switched system has an invariant set for some switching law, so we treat with the problem of design a switching law.

In the next section we describe the problem of invariant set for switched systems. We focus on a class of third-order systems where we are able to solve the problem. The problem is solved in Section 3, to accomplish this we reduce the problem to a simplified case. In this simplified case we give an invariant set. Finally we transform the invariant set to the origianl system. In Section 4 we give an example and illustrate how to calculate the invariant set. In the last section we provide the conclusions.

2 Problem description

A switched linear system is a system with several subsystems

$$\dot{x} = A_k x, \qquad k = 1, \dots, M, \tag{1}$$

where M is the number of subsystems and $A_1, \dots, A_M$ are $n \times n$ matrices. Only one subsystem is active each time, the decision of which subsystem is active is given by a switching law. A switching law says when the active subsystem changes, this decision of switching to other subsystem can be given in terms of time or state. For our purposes, the switching law depends on the state. Then, a switching law $\sigma : \mathbb{R}^n \to \{1, \dots, M\}$. Given a switching law σ the system is

$$\dot{x} = A_{\sigma(x)} x, \tag{2}$$

and we denote $\phi(\cdot; x_0, \sigma)$ the solution of (2) with the initial condition $x(0) = x_0 \in \mathbb{R}^n$.

The following definition is the fundamental property of this paper.

Definition 2.1 *A set $S \subset \mathbb{R}^n$ is an invariant set for the switched system (1) if there exist a switching law σ such that if $x_0 \in S$ then $\phi(t; x_0, \sigma) \in S$ for each $t \geq 0$.*

For second-order switched system with two subsystems, if A_1, A_2 are 2×2 matrices corresponding to each subsystem. In case that A_1 and A_2 have complex eigenvalues (two conjugate complex and non-real eigenvalues) and turn clockwise and counter-clockwise, respectively, then it is easy to check that every cone in $\mathbb{R}^2$ is an invariant set, by switching on the boundary of the cone. For more details see [8].

We desire a similar behavior in third-order, but the same argumentation it does not work for third-order. Although an invariant set can be given for a certain class in third-order.

We consider a third-order switched systems with three subsystems

$$\dot{x} = A_k x, \qquad k = 1, 2, 3, \tag{3}$$

 ISBN: 978-84-608-6082-2

where A_1, A_2, A_3 are 3×3 matrices. We are interested in switched systems with the following assumptions.

Assumption 2.2 *Each A_k, $k = 1, 2, 3$, has complex eigenvalues,i.e. the numbers $\lambda_k, a_k + b_k i, a_k - b_k i$ are the eigenvalues of A_k with $b_k \neq 0$.*

Assumption 2.3 *Let $v_k \in \mathbb{R}^3$ be an eigenvector of A_k associated to the real eigenvalue λ_k, $k = 1, 2, 3$, i.e. v_k is a non-zero vector such that $A_k v_k = \lambda_k v_k$, then v_1, v_2, v_3 are linear independent vectors.*

Our goal is to show that there is an invariant set for every switched systems with Assumption 2.2 and 2.3.

3 Invariant set for third-order switched systems

We deal with our problem in several steps. First we show the relation between the class of third-order switched system above-named and other simplified class of switched system. Then we give an invariant set for this new simplified switched systems. Finally we transform the invariant set for simplified switched system to the invariant set for a general class of switched systems.

3.1 Simplification of the class of third-order switched systems

The following proposition is useful for getting the simplified switched systems,

Proposition 3.1 *Let P be a $n \times n$ non-singular matrix. The following statements are equivalent*

1. *$P(S) \subset \mathbb{R}^n$ is an invariant set for the switched system*

$$\dot{x} = A_\sigma x,$$

2. *$S \subset \mathbb{R}^n$ is an invariant set for the switched system*

$$\dot{y} = P^{-1} A_\sigma P y.$$

Where we denote $P(S) = \{Py : y \in S\}$.

Let $v_1, v_2, v_3 \in \mathbb{R}^3$ be vectors as Assumption 2.3. We define the non-singular matrix

$$P = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}. \tag{4}$$

 ISBN: 978-84-608-6082-2

Then the matrix $P^{-1}A_kP$ has eigenvalue λ_k with eigenvector e_k, for $k = 1, 2, 3$, where $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$. By Proposition 3.1 we can replace Assumption 2.3 by the following

Assumption 3.2 *For each $k = 1, 2, 3$ the vector e_k is an eigenvector of A_k associated to the eigenvalue λ_k.*

Therefore Proposition 3.1 states that studying invariant set for switched systems (3) with Assumption 2.2 and 2.3 is equivalent for switched systems (3) with Assumption 2.2 and 3.2. Note that Assumption 3.2 is a particular case of Assumption 2.3.

3.2 Invariant octant for simplified case

In this step we are going to show that for switched system (3) with Assumption 2.2 and 3.2 there is an invariant octant.

Every octant of $\mathbb{R}^3$ is identified with three signs, i.e. each $(a, b, c) \in \{-1, +1\}^3$ is identified with the octant $O(a, b, c) = \{(x_1, x_2, x_3) : ax_1 \geq 0, bx_2 \geq 0, cx_3 \geq 0\}$.

We only consider switching laws such that the switch to other subsystem taken place in the faces of the octant. All the faces of the octant are the following

$$\begin{aligned} O(0, b, c) &= \{(0, x_2, x_3) : bx_2 > 0, cx_3 > 0\} & O(0, 0, c) &= \{(0, 0, x_3) : cx_3 > 0\} \\ O(a, 0, c) &= \{(x_1, 0, x_3) : ax_1 > 0, cx_3 > 0\} & O(0, b, 0) &= \{(0, x_2, 0) : bx_2 > 0\} \\ O(a, b, 0) &= \{(x_1, x_2, 0) : ax_1 > 0, bx_2 > 0\} & O(a, 0, 0) &= \{(x_1, 0, 0) : ax_1 > 0\} \end{aligned}$$

A switching law of this type will be called a face switching law. A face switching law σ is given by

$$\sigma(x) = \begin{cases} s_1 & \text{if } x \in O(0, b, c) \\ s_2 & \text{if } x \in O(a, 0, c) \\ s_3 & \text{if } x \in O(a, b, 0) \\ s_{12} & \text{if } x \in O(0, 0, c) \\ s_{13} & \text{if } x \in O(0, b, 0) \\ s_{23} & \text{if } x \in O(a, 0, 0) \end{cases} \tag{5}$$

where $s_1, s_2, s_3 \in \{1, 2, 3\}$ and $s_{ij} = s_i$ or $s_{ij} = s_j$ for each $1 \leq i < j \leq 3$.

The next proposition gives us a characterization for a switched system (3) with a face switching law (5) such that $O(a, b, c)$ is invariant.

Proposition 3.3 *Let $O(a, b, c)$ be an octant in $\mathbb{R}^3$ with $(a, b, c) \in \{-1, +1\}^3$, then the following statements are equivalent*

1. *$O(a, b, c)$ is an invariant set for switched system (3) with the face switching law (5),*
2. *the following statements hold*

 ISBN: 978-84-608-6082-2

(a) $ab\,e_1'A_{s_1}e_2 \geq 0$ *and* $ac\,e_1'A_{s_1}e_3 \geq 0$,
(b) $ab\,e_2'A_{s_2}e_1 \geq 0$ *and* $bc\,e_2'A_{s_2}e_3 \geq 0$,
(c) $ac\,e_3'A_{s_3}e_1 \geq 0$ *and* $bc\,e_3'A_{s_3}e_2 \geq 0$,
(d) $bc\,e_2'A_{s_1}e_3 \geq 0$ *or* $ac\,e_1'A_{s_2}e_3 \geq 0$,
(e) $ab\,e_1'A_{s_3}e_2 \geq 0$ *or* $bc\,e_3'A_{s_1}e_2 \geq 0$,
(f) $ab\,e_2'A_{s_3}e_1 \geq 0$ *or* $ac\,e_3'A_{s_2}e_1 \geq 0$.

Where $'$ *denotes transpose.*

We are going to show that there is an invariant octant for the switched system (3) with Assumption 2.2 and 3.2, to accomplish this we are going to show that there exist a face switching law and signs $(a,b,c) \in \{-1,+1\}^3$ verifying the condition given in Proposition 3.3.

Let A_1, A_2, A_3 be 3×3 matrices verifying Assumptions 2.2 and 3.2, then they are written as

$$A_1 = \begin{pmatrix} \lambda_1 & a_{12}^1 & a_{13}^1 \\ 0 & a_{22}^1 & a_{23}^1 \\ 0 & a_{32}^1 & a_{33}^1 \end{pmatrix} \qquad A_2 = \begin{pmatrix} a_{11}^2 & 0 & a_{13}^2 \\ a_{21}^2 & \lambda_2 & a_{23}^2 \\ a_{31}^2 & 0 & a_{33}^2 \end{pmatrix} \qquad A_3 = \begin{pmatrix} a_{11}^3 & a_{12}^3 & 0 \\ a_{21}^3 & a_{22}^3 & 0 \\ a_{31}^3 & a_{32}^3 & \lambda_3 \end{pmatrix} \tag{6}$$

with $a_{23}^1 a_{32}^1 < 0$, $a_{13}^2 a_{31}^2 < 0$ and $a_{12}^3 a_{21}^3 < 0$.

In order to get an invariant octant we propose the following two face switching law

1. σ_1 given by expression (5) with $s_1 = 3$, $s_2 = 1$, $s_3 = 2$, $s_{12} = 3$, $s_{13} = 2$ and $s_{23} = 1$,
2. σ_2 given by expression (5) with $s_1 = 2$, $s_2 = 3$, $s_3 = 1$, $s_{12} = 3$, $s_{13} = 2$ and $s_{23} = 1$.

Let $O(a,b,c)$ be an octant, with $(a,b,c) \in \{-1,+1\}^3$, then $O(a,b,c)$ is an invariant set for the switched system (3) with the face switching law σ_1 if, and only if, the following statements hold

1. $ab\,e_1'A_3e_2 \geq 0$ and $ac\,e_1'A_3e_3 \geq 0$,
2. $ab\,e_2'A_1e_1 \geq 0$ and $bc\,e_2'A_1e_3 \geq 0$,
3. $ac\,e_3'A_2e_1 \geq 0$ and $bc\,e_3'A_2e_2 \geq 0$,
4. $bc\,e_2'A_3e_3 \geq 0$ or $ac\,e_1'A_1e_3 \geq 0$,
5. $ab\,e_1'A_2e_2 \geq 0$ or $bc\,e_3'A_3e_2 \geq 0$,
6. $ab\,e_2'A_2e_1 \geq 0$ or $ac\,e_3'A_1e_1 \geq 0$.

Since $a_{13}^3 = a_{21}^1 = a_{32}^2 = a_{23}^3 = a_{12}^2 = a_{31}^1 = 0$, the above condition is equivalent to

$$ab\,a_{12}^3 \geq 0 \qquad bc\,a_{23}^1 \geq 0 \qquad ac\,a_{31}^2 \geq 0. \tag{7}$$

In conclusion, since $a_{12}^3 \neq 0$, $a_{23}^1 \neq 0$ and $a_{31}^2 \neq 0$, the following statements are equivalent

1. $O(a,b,c)$ is invariant for the switched system (3) with face switching law σ_1,
2. a_{12}^3 has same sign as ab, a_{23}^1 has same sign as bc and a_{31}^2 has same sign as ac.

			σ_1			σ_2		
a_{23}^1	a_{31}^2	a_{12}^3	bc	ac	ab	bc	ac	ab
$+$	$+$	$+$	$+$	$+$	$+$	$-$	$-$	$-$
$+$	$+$	$-$	$+$	$+$	$-$	$-$	$-$	$+$
$+$	$-$	$+$	$+$	$-$	$+$	$-$	$+$	$-$
$+$	$-$	$-$	$+$	$-$	$-$	$-$	$+$	$+$
$-$	$+$	$+$	$-$	$+$	$+$	$+$	$-$	$-$
$-$	$+$	$-$	$-$	$+$	$-$	$+$	$-$	$+$
$-$	$-$	$+$	$-$	$-$	$+$	$+$	$+$	$-$
$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$	$+$

Table 1: Signs of bc, ac and ab deduce from signs of a_{23}^1, a_{31}^2 and a_{12}^3, for each law σ_1 and σ_2.

By a similar argumentation and since $a_{23}^1 a_{32}^1 < 0$, $a_{13}^2 a_{31}^2 < 0$ and $a_{12}^3 a_{21}^3 < 0$, the following statements are equivalent

1. $O(a,b,c)$ is invariant for the switched system (3) with face switching law σ_2,
2. a_{12}^3 has opposite sign as ab, a_{23}^1 has opposite sign as bc and a_{31}^2 has opposite sign as ac.

We study the eight different case of sign among the numbers a_{23}^1, a_{31}^2 and a_{12}^3, and we deduce the sign of bc, ac and ab in the two case of law σ_1 or σ_2 for each above condition. This is performed in Table 1.

Finally, in Table 2 is shown all posibilities of signs of bc, ac and ab for each octant $O(a,b,c)$ with $(a,b,c) \in \{-1,+1\}^3$. We note that for all possible signs of a_{23}^1, a_{31}^2 and a_{12}^3

a	b	c	bc	ac	ab
$+$	$+$	$+$	$+$	$+$	$+$
$+$	$+$	$-$	$-$	$-$	$+$
$+$	$-$	$+$	$-$	$+$	$-$
$+$	$-$	$-$	$+$	$-$	$-$
$-$	$+$	$+$	$+$	$-$	$-$
$-$	$+$	$-$	$-$	$+$	$-$
$-$	$-$	$+$	$-$	$-$	$+$
$-$	$-$	$-$	$+$	$+$	$+$

Table 2: All posibilities of signs of bc, ac and ab for each octant $O(a,b,c)$.

we can choose $(a,b,c) \in \{-1,+1\}^3$ such that either the column of law σ_1 or the column of law σ_2 in Table 1 agrees with the signs of bc, ac and ab. In other words, for each switched system (3) with Assumption 2.2 and 3.2 there exists an octant $O(a,b,c)$, with $(a,b,c) \in \{-1,+1\}^3$, which is invariant for the switched system with either the switching law σ_1 or σ_2.

3.3 Invariant set for third-order switched systems

Consider a third-order switched system (3) such that Assumption 2.2 and 2.3 hold. If P is defined as (4), then Assumption 2.2 and 3.2 hold for switched system with subsystems $P^{-1}A_1P$, $P^{-1}A_2P$ and $P^{-1}A_3P$. As a consequence of Subsection 3.2, the last switched system has an invariant octant $O(a,b,c)$. Proposition 3.1 claims that $P(O(a,b,c))$ is an invariant set for (3), hence $P(O(a,b,c)) = \{x_1v_1 + x_2v_2 + x_3v_3 : ax_1 \geq 0, bx_2 \geq 0, cx_3 \geq 0\}$. Note that the invariant set $P(O(a,b,c))$ is an polyhedral cone and the swtiching is given on the faces.

4 Example

In this section we show an example in order to illustrate how to calculate the invariant set. Consider the switched system (3) with matrices

$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & -\frac{10}{3} \\ 0 & \frac{4}{3} & \frac{5}{3} \end{pmatrix}, \qquad A_2 = \begin{pmatrix} -3 & 2 & 2 \\ 0 & -1 & 0 \\ -4 & 4 & 1 \end{pmatrix}, \qquad A_3 = \begin{pmatrix} 4 & -2 & 0 \\ 4 & 0 & 0 \\ 1 & -\frac{3}{2} & 1 \end{pmatrix}. \tag{8}$$

Assumption 2.2 and 2.3 hold for this switched systems. Following the above notation, we calculate the vectors $v_1 = (1,0,0)$, $v_2 = (1,1,0)$ and $v_3 = (0,0,-1)$, and define the matrix

$$P = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{9}$$

Then the matrices of the simplified switched system are

$$P^{-1}A_1P = \begin{pmatrix} 1 & \frac{2}{3} & -\frac{10}{3} \\ 0 & \frac{1}{3} & \frac{10}{3} \\ 0 & -\frac{4}{3} & \frac{5}{3} \end{pmatrix},$$

$$P^{-1}A_2P = \begin{pmatrix} -3 & 0 & -2 \\ 0 & -1 & 0 \\ 4 & 0 & 1 \end{pmatrix},$$

$$P^{-1}A_3P = \begin{pmatrix} 0 & -2 & 0 \\ 4 & 4 & 0 \\ -1 & \frac{1}{2} & 1 \end{pmatrix}.$$

 ISBN: 978-84-608-6082-2

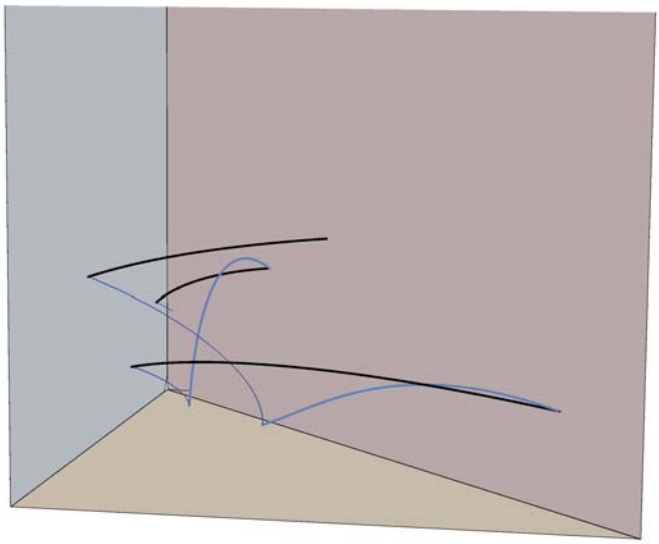


Figure 1: An example of trajectory inside the invariant set

The next step is calculating an invariant octant for the simplified switched system, to accomplish this task we get the signs of $(P^{-1}A_1P)_{23}$, $(P^{-1}A_2P)_{31}$ and $(P^{-1}A_3P)_{12}$ which are $+$,$+$ and $-$, respectively. Now, we search in Table 1 the row with signs $+,+,-$, which is the second row. In this case, it works the swtiching law σ_2 and bc, ac and ab have signs $-$, $-$ and $+$, respectively. By looking at Table 2 we deduce that the octants which agree with this signs are $(a,b,c) = (+,+,-)$ and $(a,b,c) = (-,-,+)$. If we consider $(a,b,c) = (+,+,-)$ we obtain the invariant polyhedral cone

$$P(O(+,+,-)) = \{x_1v_1 + x_2v_2 - x_3v_3 : x_1, x_2, x_3 \geq 0\} = \{(x_1 + x_2, x_2, x_3) : x_1, x_2, x_3 \geq 0\}.$$

We show in Figure 4 the invariant set and a trajectory of this switching system.

5 Conclusions

For a class of third-order switched systems we have provided an invariant set which is an polyhedral cone. Furthermore, the invariant set is calculated by a procedure.

A future work is to investigate the behavior of the switched systems under face switching laws in order to get other properties.

References

[1] H. Lin and P.J. Antsaklis, *Stability and stabilizability of switched linear systems: A survey of recent results*, IEEE Transactions on Automatic Control **54 (2)** (2009) 308–322.

[2] J. Daafouz, P. Riedinger and C. Iung, *Stability analysis and control synthesis for switched systems: A switched Lyapunov function approach*, IEEE Transactions on Automatic Control **47 (11)** (2002) 1883–1887.

[3] Z. Sun, S.S. Ge, T.H. Lee, *Controllability and reachability criteria for switched linear systems*, Automatica **38 (5)** (2002) 775–786.

[4] S.C. Bengea, R.A. DeCarlo, *Optimal control of switching systems*, Automatica **41 (1)** (2005) 11–27.

[5] Z. Sun and S.S. Ge, *Switched Linear Systems: Control and Design*, London: Springer-Verlag, 2005.

[6] N.P. Bhatia and G.P. Szeg, *Dynamical Systems: Stability Theory and Applications*, Lecture Notes in Mathematics Volume 35 1967.

[7] P. Nilsson, U. Boscain, M. Sigalotti, J. Newling, *Invariant sets of defocused switched systems*, Proceedings of the IEEE Conference on Decision and Control **6760834** (2013) 5987–5992.

[8] C. Prez and F. Bentez, *Switched convergence of second-order switched nonlinear systems*, International Journal of Control, Automation and Systems **10 (5)** (2012) 920–930.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The influence of size and temperature on the shape preference of nanoclusters and the implications for heterogeneous catalysis

Anna L. Garden[1], Andreas Pedersen[2] and Hannes Jónsson[3,4]

[1] *Department of Chemistry, University of Otago, P. O. Box 56, Dunedin, New Zealand*

[2] *Integrated Systems Laboratory, ETH Zurich, 8092 Zurich, Switzerland*

[3] *Science Institute and Faculty of Physical Sciences, University of Iceland, VR-III, Reykjavík 107, Iceland*

[4] *Department of Applied Physics, Aalto University, FI-00076, Espoo, Finland*

emails: anna.garden@otago.ac.nz, ,

Abstract

Metal nanoclusters are frequently utilised as hetergeneous catalysts due to their high catalytic activity. This high activity is due to the high surface area to volume ratio, large number of undercoordinated atoms and unique electronic properties compared with the bulk material. The catalytic activity is highly dependent on the size and specific shape of the nanocluster. Thus it follows that to have a thorough understanding of the catalytic activity and to design optimal catalysts, detailed understanding is required of a detailed understanding of the relationship between size and shape of nanoclusters is necessary, in addition to knowledge of the activity of various active sites.

In the present work, firstly the relationship between shape and activity of metal nanoclusters will be discussed. Examples will be given for H_2 evolution on Pt nanoclusters [1] and N_2 dissociation on Ru nanoclusters [2, 3]. Secondly, the influence of size and temperature on the relative stability of icosahedral, FCC and decahedral Au nanoclusters between 600-4000 atoms will be presented and dicussed in the context of rational catalyst design [4].

Key words: nanoclusters, density functional theory, heterogeneous catalysis
MSC 2000: 92E99

References

[1] E. Skúlason, A. A. Faraj, L. Kristinsdóttir, J. Hussain, A. L. Garden, H. Jónsson, *Catalytic Activity of Pt Nano-Particles for H_2 Formation*, Top. Catal. **57** (2014) 273–281.

[2] J. Gavnholt, J. Schiøtz, *Structure and reactivity of ruthenium nanoparticles*, Phys. Rev. B **77** (2008) 035404.

[3] C. Casey-Stevens, S. Lambie, E. Skúlason, A. L. Garden, *Dissociation of N_2 on ruthenium nanoparticles*, Manuscript in preparation (2016).

[4] A. Pedersen, A. L. Garden, H. Jónsson, *Size and temperature dependence of the atomic structure of Au clusters including 100 to 4000 atoms*, Manuscript in preparation (2016).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Lie symmetries and equivalence transformations for the Barenblatt-Gilman model

T. M. Garrido[1], A. A. Kasatkin[2], M. S. Bruzón[1] and R. K. Gazizov[2]

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

[2] *Laboratory of Group Analysis of Mathematical Models in Natural and Engineering Sciences (GAMMETT), Ufa State Aviation Technical University*

emails: tamara.garrido@uca.es, alexei kasatkin@mail.ru, m.bruzon@uca.es, gazizov@mail.rb.ru

Abstract

In this paper we have considered the Barenblatt-Gilman equation which models the nonequilibrium countercurrent capillary impregnation. The equation of this model is a third-order equation and the unknown function concerns to the effective water saturation.
We have applied the classical method to get the Lie group classification depending on the unknown function and we have looked for an invariant classification via equivalence transformations. Moreover, we have obtained travelling wave solutions.

Key words: Barenblatt-Gilman equation, Lie group analysis, equivalence transformations, travelling wave solutions

1 Introduction

The theory of counterflow capillary impregnation of a porous medium has been studied extensively due to its applications in various fields such as soil science, petroleum, crystal growth and flip chip underfilling [13], [12], [9] and [2].

In [1] the physical model of the non-equilibrium effects in a simultaneous flow of two immiscible fluids in porous media is presented. The Barenblatt-Gilman equation is as follows

$$u_t = \alpha \triangle \Phi(u) + \alpha\lambda \left(\triangle \Phi(u)\right)_t \tag{1}$$

 ISBN: 978-84-608-6082-2

which unknown function Φ is the effective water saturation.

2 Classical symmetries and equivalence transformations

Naturally, the study of partial differential equations plays a vital role in the physical sciences. These equations are often non-linear and solving them requires unique and creative methods. Most well-known techniques have a common feature: they exploit symmetries.

Symmetry methods provide the user with a powerful range of tools for studying equations. The required theory and description of the method can be found in [3], [6], [7], [10] and so on.

Lie classical method is specially useful in the study of Barenblatt-Gilman equation (1) due to its arbitrary function because while we are searching for symmetries it will provide a set of special forms for the unknown function Φ where it is possible to choose.

Moreover, as symmetry calculations normally are lengthy, specially for partial differential equations, we have obtained classical symmetries of (1) using the free software wxMaxima, with its associated programme *symmgrp2009.max* made by Hereman and Huard ([4], [5]), and Maple software.

We have obtained a full classification of (1) for $\alpha,\ \lambda \neq 0$ in 6 cases:

- Case 1: Φ arbitrary function.
- Case 2: $\Phi = \mathrm{e}^u$
- Case 3: $\Phi = -3u^{-1/3}$
- Case 4: $\Phi = \dfrac{u^{\gamma+1}}{\gamma+1}$
- Case 5: $\Phi = \ln u$
- Case 6: $\Phi = u$

In addition, for $\lambda = 0$, the cases from 1 to 5 admits an additional infinitesimal generator.

By the way, we have constructed the optimal system and have derived the corresponding reduced equations obtaining similarity variables and similarity solutions. And at the end, some travelling wave solutions have been obtained.

On the other hand, we have studied the equivalence transformations. Because of the presence of an arbitrary function in the Barenblatt-Gilman equation it is interesting to obtain an equivalence classification. So applying the method suggested by Ovsyannikov [11], which main part is the use of a secondary prolongation [8], we have obtained few scaling and translating transformations.

References

[1] G. I. Barenblatt, and A. A. Gilman, *Mathematical Model of the Countercurrent Capillary Impregnation*, Journal of Engineering Physics and Thermophysics, **52(3)** (1987) 456–461.

[2] G. I. Barenblatt, T. W. Patzek and D. B.Silin, *The Mathematical Model of Nonequilibrium Effects in Water-Oil Displacement*, Society of Petroleum Engineers, **8(4)**(2003) 409–416.

[3] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Springer-Verlag, New York, 1989.

[4] W. Hereman, and B. Huard, *symmgrp2009. max: A macsyma/maxima program for the calculation of lie point symmetries of large systems of differential equations.* `http://inside.mines.edu/~whereman/`

[5] B. Champagne, W. Hereman, and P. Winternitz, *The computer calculations of Lie point symmetries of large systems of differential equations*, Computer Physics Communications, **66** (1991) 319–340.

[6] N. H. Ibragimov, *CRC handbook of Lie group analysis of differential equations, vols. 1-3*, CRC Press, Florida, 1994-1996.

[7] N. H. Ibragimov, *Elementary Lie group analysis and ordinary differential equations*, Wiley, Chichester, 1999.

[8] N. H. Ibragimov, *Equivalence groups and invariants of linear and non-linear equations*, ALGA publications, **1** (2004) 9–69.

[9] G. L. Lehmann, T. Driscoll, N. R. Guydosh, P. C. Li and E. I.Cotts, *Underflow process for direct-chip-attachment packaging*, IEEE Transactions on Components Packaging and Manufacturing Technology Part A, **21(2)** (1998) 266–274.

[10] P. J. Olver, *Applications of Lie groups to differential equations*, Springer-Verlag, Berlin, 1986.

 ISBN: 978-84-608-6082-2

[11] L. V. Ovsyannikov, *Group analysis of differential equations*, Nauka, Moscow, 1978. English Translation by Academic Press, New York, 1982.

[12] W. B. Young and W. L. Yang, *Underfill viscous flow between parallel plates and solder bumps*, IEEE Transactions on Components and Packaging Technologies, **25(4)** (2002) 695–700.

[13] W. B. Young, *Capillary impregnation into cylinder banks*, *Journal of Colloid and Interface Science*, **273** (2004) 576-580.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

From clusters to the liquid state: explaining the anomalous melting temperatures of gallium clusters

Nicola Gaston[1] and Krista G. Steenbergen[2]

[1] *The MacDiarmid Institute for Advanced Materials and Nanotechnology,, Department of Physics, The University of Auckland, New Zealand*

[2] *Centre for Theoretical Chemistry and Physics, Massey University, New Zealand*

emails: n.gaston@auckland.ac.nz, kgsteen@gmail.com

Abstract

Elemental gallium is a molecular metal, a phenomenon which partly explains its low melting temperature. Small clusters of gallium, in contrast, have been shown to melt at much higher temperatures than the bulk metal. We complete an in-depth structural analysis at finite temperatures, based on the results of first-principles Born-Oppenheimer molecular dynamics simulations. We find that the anomalous melting temperatures can only be explained through an analysis of the nature of the liquid state in the clusters. *Key words: gallium, melting, polymorphism, phase transitions*

1 Introduction

The low melting temperature of elemental gallium is an anomaly of significant interest for our understanding of metallic systems. Called a 'molecular metal' due to the coexistence of covalently-bound dimers with the metallicity of the α-phase, gallium also adopts a range of low-temperature structures β, γ, δ, and ϵ, as well as a range of structures found under pressure. The high-pressure allotropes, in contrast to the low-temperature structures, become denser and more metallic with pressure, as may be expected.

However, none of the extensive bulk phase-diagram of gallium can yet explain the discovery of greater-than-bulk melting in small clusters of gallium, at sizes ranging from 17 to 55 atoms[1, 2, 3]. Numerous experimental and theoretical studies have searched for the origin of this phenomenon, which contravenes the understood variation of melting temperature with size, known as melting point depression.

 ISBN: 978-84-608-6082-2

Previous studies have demonstrated that first principles Born-Oppenheimer MD simulations can accurately reproduce the experimentally observed variation of specific heat with temperature, which suggests that the questions of how these clusters melt, and why the melting temperatures are elevated, are able to be answered. However, previous studies of melting of these clusters have compared melting characteristics *only* to the global minimum structure, without fully addressing structural changes that take place at finite temperature, both below and above the melting temperature.

In our previous work [4, 5, 6] we have demonstrated the ability of density functional theory calculations to reproduce the experimental findings, in particular the strong size-sensitivity of the melting temperatures. We have also discussed in depth the rich potential energy landscape of the clusters, in particular the transitions between different structural classes [8].

2 The liquid state of gallium clusters

In this work we have analysed the liquid state of the clusters, and found that in contrast to the usual expectations of spehrical drop-like behaviour the clusters are distorted triaxially for large proportions of the simulation time. The shortest axis of the clusters is consistently a length suggestive of the maintenance of a bilayer structure, although one in which the individual atoms retain significant mobility in the two dimensional plane.

These results explain the higher melting temperatures of the clusters as being due to the lowered entropy of the liquid phase at small sizes.[9]

Acknowledgements

This work has been supported by the Marsden Fund of the Royal Society of New Zealand under contract IRL0801. We thank the New Zealand eScience Infrastructure (NeSI), particularly the BlueFern (University of Canterbury, nesi67) and Pan (University of Auckland, nesi7) supercomputer teams for computational time and support.

References

[1] G. A. Breaux, B. Cao, and M. F. Jarrold. *Second-Order Phase Transitions in Amorphous Gallium Clusters.* J. Phys. Chem. B, **109**, 16575–16578, 2005.

[2] G. A. Breaux, D. A. Hillman, C. M. Neal, R. C. Benirschke, and M. F. Jarrold. *Gallium Cluster "Magic Melters".* J. Amer. Chem. Soc., **126**, 8628–8629, 2004.

[3] G. Breaux, R. Benirschke, T. Sugai, B. Kinnear, and M. Jarrold. *Hot and Solid Gallium Clusters: Too Small to Melt.* Phys. Rev. Lett., **91**, 215508, 2003.

[4] K. G. Steenbergen and N. Gaston. *Geometrically induced melting variation in gallium clusters from first principles.* Phys. Rev. B, **88**, 161402, 2013.

[5] K. G. Steenbergen and N. Gaston. *First-principles melting of gallium clusters down to nine atoms: structural and electronic contributions to melting.* Phys. Chem. Chem. Phys., **15**, 15325, 2013.

[6] K. G. Steenbergen, D. Schebarchov, and N. Gaston. *Electronic effects on the melting of small gallium clusters.* J. Chem. Phys., **137**, 144307, 2012.

[7] K. G. Steenbergen and N. Gaston. *Two worlds collide: Image analysis methods for quantifying structural variation in cluster molecular dynamics.* J. Chem. Phys., **140**, 064102, 2014.

[8] K. G. Steenbergen and N. Gaston. *Quantum Size Effects in the Size-Temperature Phase Diagram of Gallium: Structural Characterization of Shape-Shifting Clusters* Chem. - Eur. J. **21**, 2862–2869, 2015.

[9] K. G. Steenbergen and N. Gaston. *A Two-Dimensional Liquid Structure Explains the Elevated Melting Temperatures of Gallium Nanoclusters.* Nano Lett., **16**, 21, 2016.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Solving second order non-linear elliptic partial differential equations using generalized finite difference method

Luis Gavete[1], Francisco Ureña[2], Juan J. Benito[2], Ángel García[2], Miguel Ureña[2] and Eduardo Salete[2]

[1] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid (UPM)*

[2] *Departamento de Construcciń y Fabricación, Universida Nacional de Educación a Distancia (UNED)*

emails: lu.gavete@upm.es, fuprieto@terra.com, jbenito@ind.uned.es, angelochurri@gmail.es, miguelurenya@gmail.es, esalete@ind.uned.es

Abstract

The generalized finite difference method (GFDM) has been proved to be a good procedure to solve several linear partial differential equations (pde's): wave propagation, advection-diffusion, plates, beams, etc.
The GFDM allows us to use irregular clouds of nodes that can be of interest for modelling non-linear pde's.
This paper shows the application of the GFDM to solving different non-linear problems including applications to heat transfer, acoustics and problems of mass transfer.

Key words: meshless methods, generalized finite difference method, non-linear partial differential equations, Newton-Raphson method.
MSC 2000: 65M06, 65M12

1 Introduction

Non-linear pdes govern different problems related with heat transfer, acoustic, combustion theory, diffusion, equilibrium, and some others [14, 16, 17, 18].
Modern numerical methods, in particular those for solving non-linear PDEs, have been developed in recent years using finite differences, finite elements, finite volume or spectral methods. A review of numerical methods for non-linear partial differential equations is

given by Tadmor [15]. Meshless or meshfree methods can be also used for solving non-linear partial differential equations [19]. In this paper we use a meshless method called generalizad finite difference method (GFDM) for solving different partial elliptic non-linear pdes.
Researchers as Jensen [1], Liszka and Orkisz [2], Orkisz [3] and Perrone and Kao [4] have contributed to develop the GFDM in different aspects of its applications.
Benito, Gavete and Ureña [5, 6, 7, 8] have developed the explicit formulae and h-adaptive method for the solution of the pdes in 2-D, have also developed several applications of the GFDM for fundamentally problems of elasticity [9], advection-diffusion [10] and wave propagation [11, 12, 13] in 2-D.
This paper shows that the GFDM can be applied for solving nonlinear pde's with different irregular clouds of points in 2D.
The paper is organized as follows. In section 2 the explicit formulae in the GFDM are obtained and its applications using Newton-Raphson method for solving non-linear equations elliptic partial differential are showed. In section 3, benchmark tests of several examples are showed. Sections 4, 5, 6,7 and 8 exposes the results obtained for solving different non-linear problems. Finally, in section 9, some conclusions are obtained.

2 Explicit formulae in GFDM: application to non-linear partial differential equations

Consider the following non-linear problem in the domain $D = \Omega \cap \Gamma \subset \mathbb{R}^2$

$$\begin{cases} L_\Omega[U] = f(x,y) & in \quad \Omega \\ L_\Gamma[U] = g(x,y) & in \quad \Gamma \end{cases} \tag{1}$$

where $L_\Omega[U]$ is a non-linear operator, $L_\Gamma[U]$ is a linear partial differential operator, Γ is the boundary of the domain Ω and f, g are known functions.
Let $D \subset \mathbb{R}^2$ be a domain and

$$M = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset D$$

a discretization of the domain D with N points. Each point of the discretization M is denoted as a node.
For each one of the nodes of the domain where the value of U is unknown E_s-star is defined as $E_s = \{\mathbf{x}_0; \mathbf{x}_1, \ldots, \mathbf{x}_s\} \subset M$ with the central node $\mathbf{x}_0 \in M$ and $\mathbf{x}_i (i = 1, \ldots, s) \in M$ is a set of points located in the neighborhood of $\boldsymbol{x_0}$. To select the points different criteria as four quadrants or distance can be used.
Consider an E_s-star with the central node $\mathbf{x}_0$, where (x_0, y_0) are the coordinates of the central node, (x_i, y_i) are the coordinates of the i^{th} node in the E_s-star, and $h_i = x_i - x_0$ and $k_i = y_i - y_0$.

If $U_0 = U(\mathbf{x}_0)$ is the value of the function at the central node of the star and $U_i = U(\mathbf{x}_i)$ are the function values at the rest of the nodes, with $i = 1, ..., s$, then, according to the Taylor series expansion it is known that:

$$U_i = U_0 + h_i \frac{\partial U_0}{\partial x} + k_i \frac{\partial U_0}{\partial y} + \frac{1}{2}\left(h_i^2 \frac{\partial^2 U_0}{\partial x^2} + k_i^2 \frac{\partial^2 U_0}{\partial y^2} + 2h_i k_i \frac{\partial^2 U_0}{\partial x \partial y} \right) + ..., \quad i = 1, ..., s \quad (2)$$

We define:

$$\mathbf{c_i}^T = \{h_i, k_i, \frac{h_i^2}{2}, \frac{k_i^2}{2}, h_i k_i\} \quad (3)$$

$$\boldsymbol{D}^T = \{\frac{\partial u_0}{\partial x}, \frac{\partial u_0}{\partial y}, \frac{\partial^2 u_0}{\partial x^2}, \frac{\partial^2 u_0}{\partial y^2}, \frac{\partial^2 u_0}{\partial x \partial y}\} \quad (4)$$

If in Eq.(2) the terms over the second order are ignored, an approximation of second order for the U_i function is obtained. This is indicated as u_i. It is then possible to define the function $B(u)$ as:

$$B(u) = \sum_{i=1}^{s} [(u_0 - u_i) + h_i \frac{\partial u_0}{\partial x} + k_i \frac{\partial u_0}{\partial y} + \\ + \frac{1}{2}(h_i^2 \frac{\partial^2 u_0}{\partial x^2} + k_i^2 \frac{\partial^2 u_0}{\partial y^2} + 2h_i k_i \frac{\partial^2 u_0}{\partial x \partial y})]^2 w_i^2 \quad (5)$$

where $w_i = w(h_i, k_i)$ is a positive weighting function.
Some weighting functions as potential $\frac{1}{dist^n}$ or exponential $exp(-n(dist^2))$ can be used, where $n \in \mathbb{N}$ If the norm given by Eq.(5) is minimized with respect to the partial derivatives, the following linear equation system is obtained (for each node of the discretization where u is unknown)

$$\boldsymbol{A}(h_i, k_i, w_i)\boldsymbol{D} = \boldsymbol{b}(h_i, k_i, w_i, u_0, u_i) \quad (6)$$

where

$$\boldsymbol{A} = \begin{pmatrix} h_1 & h_2 & \cdots & h_s \\ k_1 & k_2 & \cdots & k_s \\ \vdots & \vdots & \vdots & \vdots \\ h_1 k_1 & h_2 k_2 & \cdots & h_s k_s \end{pmatrix} \begin{pmatrix} \omega_1^2 & & & \\ & \omega_2^2 & & \\ & & \cdots & \\ & & & \omega_s^2 \end{pmatrix} \begin{pmatrix} h_1 & k_1 & \cdots & h_1 k_1 \\ h_2 & k_2 & \cdots & h_2 k_2 \\ \vdots & \vdots & \vdots & \vdots \\ h_s & k_s & \cdots & h_s k_s \end{pmatrix} \quad (7)$$

Remark
The matrix $\boldsymbol{A}$ *is positive definite.*
Proof
If (x_0, y_0) are the coordinates of the central node of a star and $(x_i, y_i)(i = 1, \cdots, s$ are the coordinates of the rest of the nodes of the star and are different, we calculate (h_i, k_i), with

 ISBN: 978-84-608-6082-2

$h_i = x_i - x_0$ and $k_i = y_i - y_0$. By including the basis $\{h_i, k_i, \frac{h_i^2}{2}, \frac{k_i^2}{2}, h_i k_i\}$ as columns of a matrix, we obtain

$$\boldsymbol{P} = \begin{pmatrix} h_1 & h_2 & \cdots & h_s \\ k_1 & k_2 & \cdots & k_s \\ \vdots & \vdots & \vdots & \vdots \\ h_1 k_1 & h_2 k_2 & \cdots & h_s k_s \end{pmatrix} \tag{8}$$

In matrix $\boldsymbol{P}$ the row vectors are linearly independent. In matrix $\boldsymbol{W}$

$$\boldsymbol{W} = \begin{pmatrix} \omega_1^2 & & & \\ & \omega_2^2 & & \\ & & \cdots & \\ & & & \omega_s^2 \end{pmatrix} \tag{9}$$

with $w_i^2 > 0$.
Then matrix $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{W}\boldsymbol{P}^{\boldsymbol{T}}$ is positive definite [20] and it has a unique Cholesky decomposition. The solution of system Eq(6) is unique and the solutions obtained for the derivatives are a linear combination of the function values obtained at the nodes. Then,

$$\mathbf{D} = \boldsymbol{A}^{-1}\mathbf{b} = \boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}(\mathbf{u} - u_0\mathbf{1}) \tag{10}$$

where $\mathbf{1} = \{1, 1, ..., 1\}^T$ and $\mathbf{u} = \{u_1, u_2, ..., u_s\}^T$

$$\mathbf{D} = \underbrace{\boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}\mathbf{e_1}}_{\mathbf{m_1}} u_1 + ... + \boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}\mathbf{e_s}u_s - \boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}(\mathbf{e_1} + ... + \mathbf{e_s})u_0 =$$

$$= -\mathbf{m_0}u_0 + \sum_{i=1}^{s} \mathbf{m_i}u_i \tag{11}$$

with

$$\mathbf{m_0} = \sum_{i=1}^{s} \mathbf{m_i} \tag{12}$$

where $\boldsymbol{e}_i (i = 1, \ldots, s)$ are the vectors of the canonical basis.
By including the explicit expressions for the values of the partial derivatives, Eq. (11), in Eqs. (1) a set of N non-linear equations are obtained

$$G_i[u_j] = 0, i, j = 1, ..., N \tag{13}$$

 ISBN: 978-84-608-6082-2

2.1 Approximation order

$$\boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}(\mathbf{u}-u_0\mathbf{1}) =$$
$$= \boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}\left(\begin{pmatrix} u_0+h_1\frac{\partial u_0}{\partial x}+...+\frac{1}{2}h_1^2\frac{\partial^2 u_0}{\partial x^2}+... \\ u_0+h_2\frac{\partial u_0}{\partial x}+...+\frac{1}{2}h_2^2\frac{\partial^2 u_0}{\partial x^2}+... \\ \vdots \\ u_N+h_1\frac{\partial u_0}{\partial x}+...+\frac{1}{2}h_s^2\frac{\partial^2 u_0}{\partial x^2}+... \end{pmatrix}-\begin{pmatrix} u_0 \\ u_0 \\ \vdots \\ u_0 \end{pmatrix}\right)=$$
$$= \boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}\begin{pmatrix} h_1\frac{\partial u_0}{\partial x}+...+\frac{1}{2}(h_1^2\frac{\partial^2 u_0}{\partial x^2}+... \\ h_2\frac{\partial u_0}{\partial x}+...+\frac{1}{2}(h_2^2\frac{\partial^2 u_0}{\partial x^2}+... \\ \vdots \\ h_N\frac{\partial u_0}{\partial x}+...+\frac{1}{2}(h_s^2\frac{\partial^2 u_0}{\partial x^2}+... \end{pmatrix}+\boldsymbol{A}^{-1}\boldsymbol{P}\boldsymbol{W}\begin{pmatrix} \Theta(h_i^2,k_i^2) \\ \Theta(h_i^2,k_i^2) \\ \vdots \\ \Theta(h_i^2,k_i^2) \end{pmatrix} \tag{14}$$

by replacing $\boldsymbol{P}$ and $\boldsymbol{W}$ in (14), and taking account that the following result is obtained

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{D}+\Theta(h_i^2,k_i^2)=\boldsymbol{D}+\Theta(h_i^2,k_i^2) \tag{15}$$

Thus, the approximation is of second order $\Theta(h_i^2,k_i^2)$.

Remark

Consider the Dirichlet problem:

$$\begin{cases} \nabla^2 U=f(x,y) & (x,y)\in\Omega \\ U(x,y)=g(x,y) & (x,y)\in\Gamma \end{cases} \tag{16}$$

being $U(x,y)$ an unknown function sufficiently derivable in the domain Ω, $f(x,y)$ and $g(x,y)$ are known functions and $f(x,y)$ is sufficiently derivable. Then, for case $s=8$ (scheme of nine-point and $h_i=k_i=h, w_i=\frac{1}{dist^3}$, the truncation error is of the order six.

Proof

On substituting for $u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8$ (values of the nodes of the star) in terms of u_0 (value of the central node) and expanding the Taylor's series, we obtain as approximation of Laplacian

$$\frac{4(u_1+u_3+u_5+u_7)+u_2+u_4+u_6+u_8-20u_0}{6h^2}=\nabla^2u_0+\frac{h^2}{12}\nabla^2(\nabla^2u_0)$$
$$+\frac{h^4}{360}[\nabla^4(\nabla^2u_0)+2\frac{\partial^4}{\partial x^2\partial y^2}\nabla^2u_0]+\frac{h^6}{3.8!}[3\nabla^8u_0+44(\frac{\partial^4}{\partial x^4\partial y^4}\nabla^4u_0)+38\frac{\partial^8u_0}{\partial x^4\partial y^4}] \tag{17}$$

Denote

$$F(x_0,y_0)=f(x_0,y_0)+\frac{h^2}{12}\nabla^2(f(x_0,y_0)+\frac{h^4}{360}[\nabla^4f(x_0,y_0)+2\frac{\partial^4}{\partial x^2\partial y^2}f(x_0,y_0)] \tag{18}$$

 ISBN: 978-84-608-6082-2

then $F(x_0, y_0)$ is known in Ω and taking into account Eq.(17), then, we can obtain a six-order accurate method of the form

$$\frac{4(u_1+u_3+u_5+u_7)+u_2+u_4+u_6+u_8-20u_0}{6h^2} = F(x_0, y_0) \tag{19}$$

because

$$\frac{4(u_1+u_3+u_5+u_7)+u_2+u_4+u_6+u_8-20u_0}{6h^2} - F(x_0, y_0) =$$
$$= \frac{h^6}{3.8!}[3\nabla^8 u_0 + 44(\frac{\partial^4}{\partial x^4 \partial y^4}\nabla^4 u_0) + 38\frac{\partial^8 u_0}{\partial x^4 \partial y^4}]$$

then a six-order approximation can be obtained [19, 21].

2.2 Newton-Raphson method

To solve the system of equations Eq.(13) the Newton-Raphson method is used.
Let us consider the Jacobian matrix

$$\boldsymbol{J} = \frac{\partial G_i(\boldsymbol{u})}{\partial u_j} \tag{20}$$

Newton-Raphson method is based in the convergence of the following series vectors considering a series of solutions U^k

$$\boldsymbol{U}^{k+1} = \boldsymbol{U}^k - \boldsymbol{J}^{-1}(\boldsymbol{U}^k)\boldsymbol{G}(\boldsymbol{U}^k) \tag{21}$$

The Newton-Raphson (N-R) error is defined as $\|\boldsymbol{U}^{k+1} - \boldsymbol{U}^k\|$, difference between two successive solutions and its considered in the last iteration.
To stop the algorithm we use two different criterion:

- $\|\boldsymbol{U}^{k+1} - \boldsymbol{U}^k\| \leq 10^{-4}$. The iteration scheme is limited by a given tolerance.
- The maximum number of iterations is 20.

A good initial guess U^0 of the discretization of (1) is required, in order to apply (21). As the Dirichlet condition is known, it can be used to interpolate the initial solution U^0 in the interior of the domain.

3 Benchmark tests.

To analyse the obtained approximation, the method has been applied firstly to the solution of six different non-linear partial differential equations (see table 1) with Dirichlet boundary conditions. The benchmark problems used correspond to the non-linear pde's with their solutions.

Table 1: Non-linear pde's and solutions for benchmark tests

non-linear pde's	exact solution
$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+(\frac{\partial U}{\partial x})^2+(\frac{\partial U}{\partial y})^2=e^{-2x}(I)$	$U(x,y)=e^{-x}sin(y)$
$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}=4U^2(\mathrm{II})$	$U(x,y)=((x+1)^2+(y+1)^2)^{-1}$
$\frac{\partial}{\partial x}(x\frac{\partial U}{\partial x})+\frac{\partial}{\partial y}(y^3\frac{\partial U}{\partial y})=U^2(\mathrm{III})$	$U(x,y)=\frac{2y}{1+xy}$
$\frac{\partial}{\partial x}(e^x\frac{\partial U}{\partial x})+\frac{\partial}{\partial y}(e^y\frac{\partial U}{\partial y})=U^3(\mathrm{IV})$	$U(x,y)=\frac{\sqrt{3}}{2}\frac{1}{\sqrt{e^{-x}+e^{-y}}}$
$\frac{\partial^2 U}{\partial x^2}+(U+1)\frac{\partial^2 U}{\partial y^2}+(\frac{\partial U}{\partial y})^2=0(\mathrm{V})$	$U(x,y)=(x+1)y-\frac{1}{12}(x+1)^4$
$U\frac{\partial^2 U}{\partial x^2}+U\frac{\partial^2 U}{\partial y^2}+(\frac{\partial U}{\partial x})^2+(\frac{\partial U}{\partial y})^2=2U^4(\mathrm{VI})$	$U(x,y)=\frac{1}{\sqrt{(x+1)^2+(y+1)^2}}$

In the following table 1 six non-linear pde's and the corresponding solutions are shown

A domain $\Omega=[0,1]\times[0,1]$ with regular clouds of nodes is used as initial with $64=8\times 8$, $100=10\times 10$, $144=12\times 12$, $196=14\times 14$ and $256=16\times 16$ nodes respectively.
Figures 1, 2 and 3 show the results obtained for quadratic (22) and maximum (23) error formulae that have been used to compare the exact solution with the approximated solution.

$$Quadratic \quad error=\sqrt{\sum_{i=1}^{N}(u_i-u_{exact.i})^2} \tag{22}$$

$$Maximun \quad error=\max_{i=1,..,N}(|u_i-u_{exact.i}|) \tag{23}$$

According with [19] it is important to keep clear the distinction between the convergence of Newton-Raphson method to a solution of the generalized finite difference equations, and the convergence of generalized finite difference approximation to the solution of the partial differential equation, then in this paper both errors N-R and global errors, Eqs.(22) and (23), are considered.

As it shown in figures 1, 2 and 3 we obtain a good convergence to the exact solution.

4 Non-linear heat transfer problems

In this section we study the case of stationary non-linear heat transfer problems where the conductivity is a function of the temperature U as in (24)

$$k(x,y,U)\frac{\partial^2 U}{\partial x^2}+k(x,y,U)\frac{\partial^2 U}{\partial y^2}+\frac{\partial k(x,y,U)}{\partial U}(\frac{\partial U}{\partial x})^2+\frac{\partial k(x,y,U)}{\partial U}(\frac{\partial U}{\partial y})^2=q(x,y) \tag{24}$$

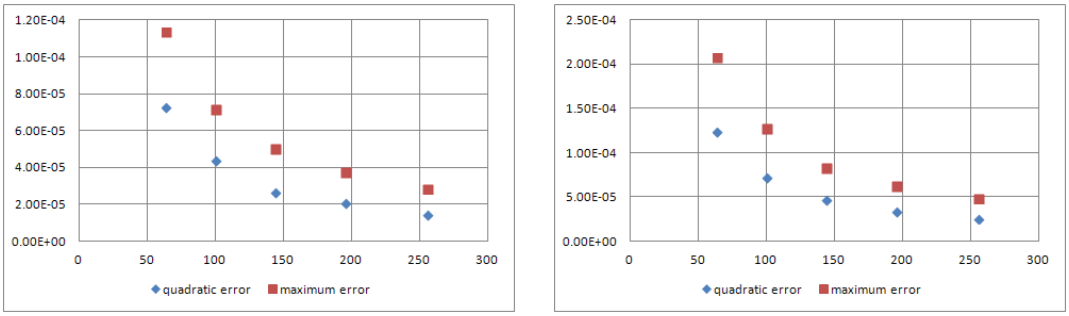


Figure 1: The quadratic and maximum error formulae solving equations (I) and (II) using clouds of nodes (with 64, 100, 144, 196 and 256)

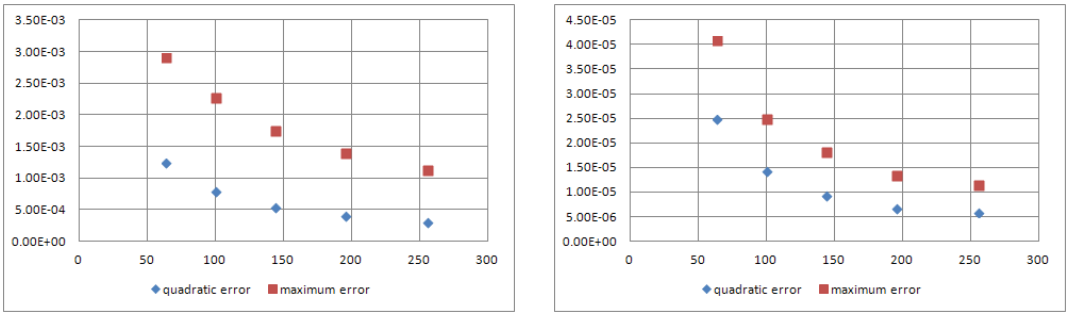


Figure 2: The quadratic and maximum error formulae solving equations (III) and (IV) using clouds of nodes (with 64, 100, 144, 196 and 256)

Different domain have been considered as shown in Fig.4.
Two different examples have been solve as particular cases of general Eq.(24)

Example 1 We analyse the non-linear pde

$$(1+U)(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}) = 4(2+x+y+x^2+y^2+xy) \tag{25}$$

It has the solution

$$1+x+y+xy+x^2+y^2 \tag{26}$$

In this example 1 the weighting function selected is $\frac{1}{dist^4}(dist = distance)$ and the criterium to select the star of nodes is the distance. The table 2 is refered to the results obtained for the three clouds of nodes (a,b,c) shown in fig.4

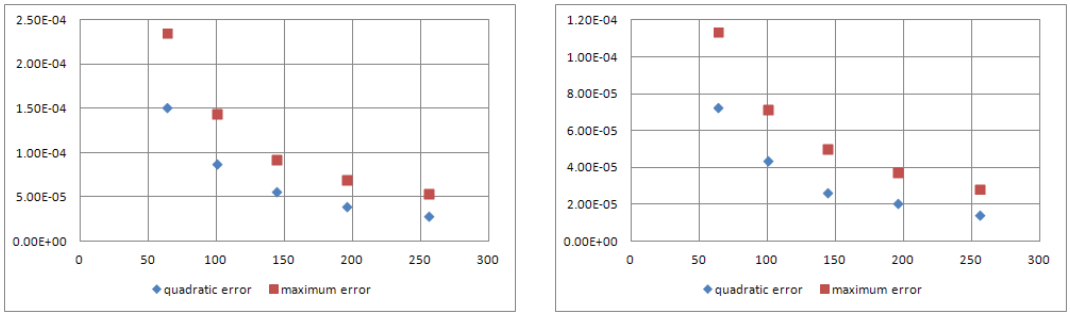


Figure 3: The quadratic and maximum error formulae solving equations (V) and (VI) using clouds of nodes (with 64, 100, 144, 196 and 256)

Table 2: N-R error, quadratic error and maximum error

Example 1	Cloud a	Cloud b	Cloud c
Iterations number	5	5	6
N-R error	$1.1014 \cdot 10^{-6}$	$9.174 \cdot 10^{-7}$	$2.556 \cdot 10^{-6}$
Quadratic error	$2.275 \cdot 10^{-7}$	$2.8193 \cdot 10^{-7}$	$2.0919 \cdot 10^{-7}$
Maximum error	$5.0333 \cdot 10^{-6}$	$7.615 \cdot 10^{-6}$	$6.909 \cdot 10^{-6}$

Example 2 Let us consider the nonlinear pde

$$(1+U)(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2})+(\frac{\partial U}{\partial x})^2+(\frac{\partial U}{\partial y})^2 = 10+10x+10y+9x^2+9y^2+12xy \quad (27)$$

It has the solution (26).
In this example 2 the weighting function selected is exponential and the criterium to select the star of nodes is the four quadrants. The table 3 is referred to the results obtained for the three clouds of nodes (a,b,c) shown in fig.4

In both examples 1 and 2, using different weighting functions and different criteria to select the star nodes, we obtain convergent and accurate results in a few iterations, the final quadratic and maximum errors obtained are very small.

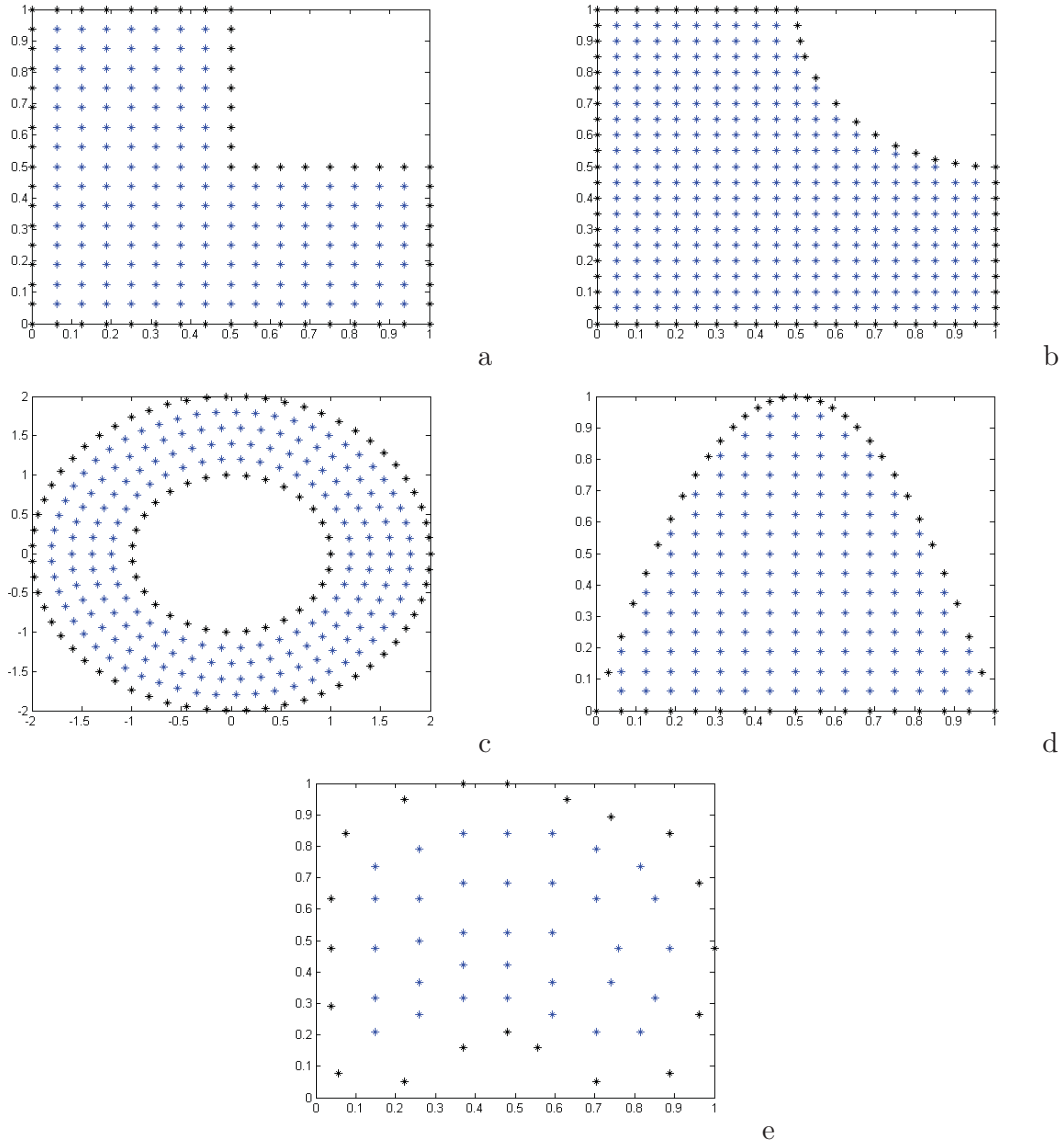


Figure 4: Irregular clouds of nodes: a) with 225 nodes, b) with 361 nodes, c) with 283 nodes, d) with 205 nodes, e) with 55 nodes

5 Heat equation with non-linear source term

We analyse a non-linear pde of combustion theory [14, 17, 18] defined as

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 6U^2 \tag{28}$$

Table 3: N-R error, quadratic error and maximum error

Example 2	Cloud a	Cloud b	Cloud c
Iterations number	5	4	6
N-R error	$2.344 \cdot 10^{-6}$	$9.254 \cdot 10^{-7}$	$4.565 \cdot 10^{-6}$
Quadratic error	$2.304 \cdot 10^{-7}$	$2.796 \cdot 10^{-7}$	$2.2889 \cdot 10^{-7}$
Maximum error	$5.580 \cdot 10^{-6}$	$8.056 \cdot 10^{-6}$	$7.480 \cdot 10^{-6}$

It has the solution

$$U = (x + \frac{y}{\sqrt{3}} + 1)^{-2} \tag{29}$$

In order to solve this problem, the criteria of the distance and the potential weighting function have been used.
The four clouds (a,b,d,e) of nodes used are shown in fig. 4.
The errors and numbers of iterations of N-R method are shown in table 4.

Table 4: N-R error, quadratic error and maximum error

PDE 28	Cloud a	Cloud b	Cloud d	Cloud e
Iterations number	3	3	3	3
N-R error	$1,1316 \cdot 10^{-5}$	$2,938 \cdot 10^{-5}$	$4,066 \cdot 10^{-5}$	$2,148 \cdot 10^{-5}$
Quadratic error	$7,306 \cdot 10^{-3}$	$8,379 \cdot 10^{-3}$	$6,418 \cdot 10^{-3}$	$7,34 \cdot 10^{-3}$
Maximum error	$1,372 \cdot 10^{-2}$	$1,512 \cdot 10^{-2}$	$1,148 \cdot 10^{-2}$	$1,134 \cdot 10^{-2}$

6 Stationary pde of Khokhlov-Zabolotskya

The stationary equation of Khokhlov-Zabolotskya [14, 16] describes different phenomena in acoustics, non-linear mechanics and mass transfer. It general form is given by

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial}{\partial y}[(\alpha U + \beta)\frac{\partial U}{\partial y}] = 0 \tag{30}$$

and the analytical solution for $\alpha = \beta = 1$, is

$$U(x, y) = (x + 1)y - \frac{1}{12}(x + 1)^4 \tag{31}$$

 ISBN: 978-84-608-6082-2

The GFDM has been used to solve eq.(30) with the criterium of the distance to select the nodes of the stars with the irregular clouds of nodes shown in fig. 4. Potential weighting function has been employed. In table 5 the results obtained are shown.

Table 5: N-R error, quadratic error and maximum error

PDE 30	Cloud a	Cloud b	Cloud c	Cloud d	Cloud e
Iterations number	6	6	10	5	5
N-R error	$3,047 \cdot 10^{-5}$	$7,435 \cdot 10^{-5}$	$1,683 \cdot 10^{-4}$	$1,940 \cdot 10^{-4}$	$1,517 \cdot 10^{-5}$
Quadratic error	$1,438 \cdot 10^{-5}$	$9,406 \cdot 10^{-6}$	$8,476 \cdot 10^{-5}$	$2,314 \cdot 10^{-4}$	$7,914 \cdot 10^{-6}$
Maximum error	$2,876 \cdot 10^{-5}$	$2,389 \cdot 10^{-5}$	$7,452 \cdot 10^{-4}$	$7,938 \cdot 10^{-4}$	$2,579 \cdot 10^{-5}$

7 Full non-linear equation

Let us consider the following full non-linear pde

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + (\frac{\partial U}{\partial x})^2 + (\frac{\partial U}{\partial y})^2 = e^{-2x} \tag{32}$$

It has the solution

$$U(x, y) = e^{-x} sin(y) \tag{33}$$

The GFDM has been used to solve Eq.(32) with the criterium of the distance to select the nodes of the stars with the irregular clouds of nodes shown in figure 4. Potential weighting function has been employed. In table 6 the results obtained are shown.

Table 6: N-R error, quadratic error and maximum error

PDE 32	Cloud a	Cloud b	Cloud c	Cloud d	Cloud e
Iterations number	3	3	3	3	3
N-R error	$3,818 \cdot 10^{-4}$	$4,787 \cdot 10^{-4}$	$3,552 \cdot 10^{-4}$	$8,615 \cdot 10^{-4}$	$1,230 \cdot 10^{-4}$
Quadratic error	$3,825 \cdot 10^{-6}$	$4,845 \cdot 10^{-6}$	$1,421 \cdot 10^{-5}$	$1,159 \cdot 10^{-4}$	$5,033 \cdot 10^{-6}$
Maximum error	$8,1241 \cdot 10^{-6}$	$1,6314 \cdot 10^{-5}$	$4,241 \cdot 10^{-5}$	$2,283 \cdot 10^{-4}$	$1,534 \cdot 10^{-5}$

8 Exponential equation of combustion theory

This type of pde's Eq.(34) are related with the combustion theory and the heat extinction [17, 18].

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = e^U \tag{34}$$

the analytical solution is

$$U(x,y) = \ln(\frac{4}{(x+y+1)^2}) \tag{35}$$

The GFDM has been used to solve Eq.(34) with the criterium of the distance to select the nodes of the stars with the irregular clouds of nodes shown in figure 4. Potential weighting function has been employed. In table 7 the results obtained are shown.

Table 7: N-R error, quadratic error and maximum error

PDE 34	Cloud a	Cloud b	Cloud c	Cloud d	Cloud e
Iterations number	3	3	3	3	3
N-R error	$1,078 \cdot 10^{-6}$	$4,561 \cdot 10^{-6}$	$4,831 \cdot 10^{-9}$	$2,348 \cdot 10^{-6}$	$9,685 \cdot 10^{-9}$
Quadratic error	$1,059 \cdot 10^{-2}$	$1,111 \cdot 10^{-2}$	$3,289 \cdot 10^{-2}$	$6,444 \cdot 10^{-3}$	$3,109 \cdot 10^{-3}$
Maximum error	$2,276 \cdot 10^{-2}$	$2,382 \cdot 10^{-2}$	$5,651 \cdot 10^{-2}$	$1,382 \cdot 10^{-2}$	$7,164 \cdot 10^{-3}$

9 Conclusions

This paper shows a scheme in generalized finite differences to the solution of the non-linear pde's using the Newton-Raphson method.
The example provided illustrates the viability of the application of GFDM for solving elliptical non-linear pde's in 2D. The efficiency of the proposed method is clearly shown.
The accuracy of the GFDM has been tested in different non-linear pde's, including different cases related with acoustics, heat transfer, mass transfer, heat extinction, combustion. Numerical results for several non-linear problems validate the use of GFDM to solve this type of problems.

10 Ackonwledgement

The authors acknowledge the support of the Escuela Técnica Superior de Ingenieros Industriales (UNED) of Spain, project 2016-IFC02.

References

[1] P. S. Jensen, Finite difference technique for variable grids, Computer& Structures 1972;2: 17-29.

[2] T. Liszka, J. Orkisz, The Finite Difference Method at Arbitrary Irregular Grids and its Application in Applied Mechanics, Computer & Structures 1980;11: 83-95.

[3] J. Orkisz, Finite Difference Method (Part, III) in handbook of Computational Solid Mechanics, M. Kleiber (Ed.), Spriger-Verlag, Berlin 1998.

[4] N. Perrone, R. Kao, A general finite difference method for arbitrary meshes, Computer& Structures 1975;5: 45-58.

[5] J. J. Benito, F. Ureña, L. Gavete, Influence several factors in the generalized finite difference method, Applied Mathematical Modeling 2001; 25: 1039-1053.

[6] J. J. Benito, F. Ureña, L. Gavete, R. Alvarez, An h-adaptive method in the generalized finite difference, Comput. Methods Appl. Mech. Eng. 2003;192:735-759.

[7] J. J. Benito, F. Ureña, L. Gavete, Leading-Edge Applied Mathematical Modelling Research (chapter 7), Nova Science Publishers, New York, 2008.

[8] J. J. Benito, F. Ureña, L. Gavete, B. Alonso, Application of the Generalized Finite Difference Method to improve the approximated solution of pdes, Computer Modelling in Engineering & Sciences 2009;38: 39-58.

[9] F. Ureña, E. Salete, J. J. Benito, L. Gavete, Solving third and fourth order partial differential equations using GFDM, Application to solve problems of plates, International Journal of Computer Mathematics 2012;89(3): 366-376

[10] F. Ureña, J. J. Benito, L. Gavete, Application of the generalized finite difference method to solve the advection-diffusion equation, Journal of Computational and Applied Mathematics 2011;235: 1849-1855.

[11] F. Ureña, J. J. Benito, E. Salete, L. Gavete, A note on the application of the generalized finite difference method to seismic wave propagation in 2-D, Journal of Computational and Applied Mathematics 2011;236(12): 3016-3025.

[12] J. J. Benito, F. Ureña, L. Gavete, B. Alonso, Solving parabolic and hyperbolic equations by Generalized Finite Difference Method, Journal of Computational and Applied Mathematics 2007;209: 208-233.

 ISBN: 978-84-608-6082-2

[13] J. J. Benito, F. Ureña, L. Gavete, E. Salete, A. Muelas, A GFDM with PML for seismic wave equations in heterogeneous media, Journal of Computational and Applied Mathematics 2013; 252: 40-51.

[14] A. D. Polyanin, V. F. Zaitsev, Handbook of nonlinear partial differential equations, 2004 by Chapman & Hall/CRC, ISBN 1-58488-355-3.

[15] A. Tadmor, A review of numerical methods for non-linear partial differential equations, Bulletin of the American Mathematical Society 2012; 42(4):507-554.

[16] E. A. Zabolotskaya, R. V. Khokhlov, Quasi-planes waves in the nonlinear acoustic of confined beams, Sov. Phys. Acoust. 1969; 15(1: 35-40.

[17] J. L. Vzquez, Domain of existence and blow-up for the exponential reaction-diffusion equation, Indiana Univ. Math. Journal 1999; 48(2):677-709.

[18] F. A. Williams, Combustion Theory 2^{nd}, Ed. Addison-Weslwy, 1985

[19] R. J. LeVeque, Finite Difference Methods for Differential Equations, University of Washington, 2004

[20] M. G. Armentano, Estimaciones de Error para aproximaciones obtenidas usando cuadrados mínimos con peso variable, Tesis doctoral, Universidad de Buenos Aires, 2000

[21] G. D. Smith, Numerical Solution of Partial Differential Equations: Finite Difference Methods 2^{nd} Edition, Oxford University Press,1978

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Using Optimal Control Theory with Mosquito Repellents and Vaccination Applied to Dengue Disease Prevention and Reduction Managment, a First Toy Study with Analytically Treatable Models

Peyman Ghaffari[1], Karunia Putra Wijaya[2], Maíra Aguiar[1], Luís Mateus[1], Thomas Götz[2] and Nico Stollenwerk[1]

[1] *Biomathematics and Statistics Group, Centro de Matemática, Aplicações Fundamentais e Investigação Operacional CMAF-CIO, Department of Mathematics, Universidade de Lisboa, Campo Grande, Lisboa, Portugal*

[2] *Mathematical Institute, University of Koblenz, 56070 Koblenz,Germany*

emails: pgsaid@fc.ul.pt, karuniaputra@uni-koblenz.de, maira@ptmat.fc.ul.pt, lgmateu@fc.ul.pt, goetz@uni-koblenz.de, nico.biomath@gmail.com

Abstract

Zika, dengue, chikungunya and yellow fever are examples of vector-borne diseases transmitted by day-time active mosquitoes. In 128 countries, in particular in tropic and sub-tropic regions of Asia and Latin America these diseases are a major health risk and a negative economic factor. In highly populated countries, like Thailand, Brazil, India and Pakistan flavivirus infections transmitted by Aedes mosquitos contribute to the high disease burden. Classical mosquito control measures, like bed-nets and municipal spraying in the streets, have proven to be of little effectiveness in combating disease cases. In mosquito control, some activities in demonstration of efficacy using bed-nets via the WHO are performed. However bed nets are not very efficient against the disease. One reason is that vectors of dengue, the species *Aedes-Egypti* and *Aedes-Albopictus*, are active in the morning and evening but not very active at night. A new generation of disease prevention is therefore required. Epidemiologists are encouraged to investigate new measures, like vaccination and mosquito repellence. In this paper we study a toy-model which mimics the vaccination and repellency factor in the linear infection model. Numerical analysis with optimal control theory is also performed.

Key words: optimal control theory, repellency, dengue fever, vaccination, mosquito reduction management, linear infection model

1 Introduction

Zika infection, dengue fever, chikungunya and yellow fever are examples of vector-borne diseases transmitted by day-time active mosquitoes. In 128 countries, in particular in tropic and sub-tropic regions of Asia and Latin America these diseases are a major health risk and a negative economic factor.

In recent years, however, vector-borne diseases and especially dengue fever are occurring in Europe. Some reasons for this are the worldwide flow of trade and travelling tourism. Increasing urbanization, as well as regional warming due to global climate change, have amplified the spread of mosquitoes like Aedes albopictus in Europe. In 2010, infections with the dengue virus were registered in Croatia, France and Italy. In 2012 and 2013, more than 2000 autochtonous patients having dengue fever were found on the isle of Madeira/Portugal transmitted by *Aedes aegypti*. Chikungunya infections occurred in Italy (2007) and Spain/France (2015). Eggs, larvae, pupae and adult mosquitoes of Aedes albopictus were repeatedly detected in the south of Germany in autumn 2014 and in 2015. Researchers assume from these findings that Asian tiger mosquitoes can survive the winter and settle in Germany.

Over the past few decades, the incidence of dengue has grown dramatically. Recent studies indicate the existence of approximately 390 million dengue infections per year and that 3,9 billion people, in 128 countries including Thailand, Brazil, India and Pakistan, are at risk of being infected with the dengue virus. The WHO has set the goal to constrain and control the spreading of dengue fever by 2020, however there are major obstacles in achieving this goal. Some vaccines are in advanced trial stages, but not effective against all serotypes, with the Phase 3 results of the Sanofi Pasteur vaccine as front runner just concluded, and have negative effects in some age classes. WHO guidelines for vaccine trials are very detailed and specific in their requirements of scientific investigation before licensing, with phases 1, 2, 2b and 3, and finally phase 4 after licensing.

As already mentioned, for dengue fever first vaccine trials are running, but the results are not satisfactory. In general regarding mosquito vector-borne diseases vaccines are quite imperfect like DenVaxia for dengue fever, recently licenced by Sanofi-Pasteur, or vaccines do not yet exist as is the case for the zika virus. In relation to yellow fever the vaccine is even in some cases lethal. Classical mosquito control measures, like bed-nets and municipal spraying in the streets, have proven to be of little effectiveness in combating disease cases. In mosquito control, some activities in demonstration of efficacy using bed-nets via the WHO are performed. However bed nets are not very efficient against the disease. One reason is that vectors of dengue, the species *Aedes-Egypti* and *Aedes-Albopictus* are active in the morning and evening, but not very active at night. Another important aspect in eliminating mosquitoes by classical pesticides and insecticides, beside the danger to human health, is that the elimination of mosquitoes, would also deprive many fish, birds, and reptiles of a food source and even destroy critical pollinator for plants.

 ISBN: 978-84-608-6082-2

In future research we will investigate SIR-type models with repellency and vaccination and analyse with optimal control theory. Here we first study a toy model which can in many aspects be treated analytically, and can already capture some simplest aspects of repellents respectively vaccination. Then numerical methods are studied to relax the need for analyticity in the models. In order to calculate numerically the influence of repellency and vaccination in the model we will use the gradient method and shooting method and compare the results. For some aspects see also [**?**, **?**]. Ones the effectiveness of different control measures in known, like e.g. in the case of the dengue vaccine [**?**] using the experimental data obtained during the phase 3 trial, e.g. [**?**] or like new generations of mosquito repellent applications (including nano-particles in cloths etc.) the next step, of course, is the suggestion of optimal strategies to combat vector-borne diseases like dengue fever, and it would be a matter of optimal control as a mathematical field.

2 The mathematical model for dengue fever

2.1 The general SIRUV-model for coupling mosquito to human epidemiology including repellency and vaccination

For pure human disease epidemiology, we assume the usual SIR model given by the ODEs:

$$\begin{aligned} \frac{d}{dt}S &= \mu(N-S) - \frac{\beta}{N}SI \\ \frac{d}{dt}I &= \frac{\beta}{N}SI - (\gamma+\mu)I \\ \frac{d}{dt}R &= \gamma I - \mu R \end{aligned} \tag{1}$$

with state variables S for susceptible humans, I for infected and R for recovered humans. The population of humans $N = S + I + R$ is assumed constant. The infection rate is given by β, recovery rate γ and birth and death rate for humans by μ.

The stationary states are easily otained. For the human infection we obtain the trivial disease-free equilibrium stationary state $I_1^* = 0$ and the non-trivial case $I_2^* = (\mu/(\gamma+\mu)(1-(\gamma+\mu)/\beta)N$. Respectly we have the two stationary states for the susceptibles $S_1^* = N$ and $S_2^* = ((\gamma+\mu)/\beta)N$ and for the recovered in both cases $R^* = N - S^* - I^*$.

Now we add to the ODEs (1) the susceptible mosquito population U and the infected mosquitos V. In the easiest case the total population size M adds up to $M = U + V$. We assume since mosquitos do not have an immune system they cannot recover from the infection. So the resulting ordinary differential equations give the following system:

$$\frac{d}{dt}S = \mu(N-S) - \frac{\beta}{M}SV - \upsilon S$$

 ISBN: 978-84-608-6082-2

$$\begin{aligned}
\frac{d}{dt}I &= \frac{\beta}{M}SV - (\gamma+\mu)I \\
\frac{d}{dt}R &= \gamma I - \mu R + \upsilon S \\
\frac{d}{dt}U &= \psi - \nu U - \frac{\vartheta}{N}UI \\
\frac{d}{dt}V &= \frac{\vartheta}{N}UI - \nu V
\end{aligned} \tag{2}$$

with mosquito birth rate ψ, infection rate from human to mosquito ϑ and mosquito mortality ν, mosquito population size $M = U(t) + V(t)$ assumed constant, hence for now $\psi := \nu \cdot M$. We also included in the above equations (2) the vaccination factor υ. Here the repellent acts as reducing contact rates between humans and mosquitoes, hence β and ϑ, where in time scale separation or center manifold analysis the simplest version already gives $\beta \cdot \vartheta$ as contact parameter in the effective SIR model .

2.2 The simple case of linear infection model and optimal control

Using the described model in the last section we use the simple case of the linear infection model to understand the application of the optimal control method in these above mentioned models. The linear infection model with reaction scheme

$$S + I^* \quad \xrightarrow{\beta} \quad I + I^* \tag{3}$$

with susceptibles S meeting outside infected I^* assumed in equilibrium with infection rate β brings the different aspects of of modelling and analysing disease control measures together, since

$$\begin{aligned}
S_c + I^* &\xrightarrow{\beta} I_c + I^* \\
S_v + I^* &\xrightarrow{(1-k_v)\beta} I_v + I^*
\end{aligned}$$

is used in its stochastic version to describe vaccine trials . Here k_v is the vaccine efficacy which can be estimated e.g. in a Bayesian framework from empirically given numbers of infected in the vaccine group I_v, ones β is estimated from the number of infected in the control group I_c. And S_c and S_c are given by the group sizes of control group and vaccine group. Vaccines, however, are not always completely effective or even in some age groups like the dengue vaccine dangerous with occasionally negative vaccine efficacy $k_v < 0$, giving quite interesting effects in more complex multi-strain modelles including such imperfect vaccines in itself .

Hence, other control measures like mosquito repellents, modelled as

$$\begin{aligned}
S_c + I^* &\xrightarrow{\beta} I_c + I^* \\
S_M + I^* &\xrightarrow{(1-k_M)\beta} I_M + I^*
\end{aligned}$$

should be studied, with an as yet to investigate mosquito repellent efficacy k_M.

The reaction scheme gives the rate equation

$$\frac{d}{dt}I = \frac{\beta}{N}SI^* = \beta^* S \tag{4}$$

with $\beta^* = (\beta/N)I^*$ and then omiting the star, but keeping conservation of population size N constant, hence

$$\frac{d}{dt}I = \beta S = \beta(N - I) \tag{5}$$

defining the linear infection model . For the optimal control problem the linear infection model is given by

$$\frac{d}{dt}I = \beta_0(N - I) \tag{6}$$

with natural infection rate β_0, as for example measured from a control group study. For any control measure, vaccination of mosquito control, we obtain then

$$\frac{d}{dt}I = (1 - k)\beta_0(N - I) = (\beta_0 - k\beta_0)(N - I) \tag{7}$$

with a maximally possible control $u_0 := k\beta_0$ for a vaccination of mosquito repellency group with all members vaccinated (hence 100% coverage). Any actual control strategy $u(t)$ then should be larger zero and smaller than the maximum, hence $0 < u(t) < u_0$.

We introduce the control signal $u(t)$ in $\beta(t) = (\beta_0 - u(t))$ in the above optimal control problem (**??**) with the simple solution of the ODE

$$\begin{aligned} I(t) &= N - (N - I_0)e^{-\int_{t_0}^{t}\beta(\tilde{t})d\tilde{t}} \\ &= N - (N - I_0)e^{-\int_{t_0}^{t}\{\beta_0 - u(\tilde{t})\}d\tilde{t}} \qquad (8) \\ &=: I[u(\tilde{t})] \qquad (9) \end{aligned}$$

and then the optimal control problem minimizes the cost function

$$\mathcal{J} = \int_{t_0}^{T} \{\frac{1}{2}kI^2 + \frac{1}{2}\ell u^2\}\, dt \tag{10}$$

By using the variation of the cost function

$$\frac{\delta\mathcal{J}[u(t)]}{\delta u(t)} = 0 \tag{11}$$

and for numerical analysis discretisation we can then solve the above problem. For the solution we discretize the total time T into n steps with size Δt

$$T = n\Delta t \tag{12}$$

 ISBN: 978-84-608-6082-2

and hence the continuous $I(t)$ becomes time discretisize time using Eq. (**??**)

$$I_i = N - (N - I_0)e^{-\sum_{k=0}^{i-1}\{\beta_0 - u_k\}\,\Delta t} \tag{13}$$

and for convenience $B_i := e^{-\sum_{k=0}^{i-1}\{\beta_0 - u_k\}\,\Delta t}$. Note that we run the sum from $k = 0$ to $k = i - 1$. The cost function becomes plugging the above into Eq. (13), and using $m := \ell/k$ with unimportant factor k in the cost function

$$\begin{aligned} \mathcal{J}(T) &= \sum_{i=1}^{n}\{\frac{1}{2}I_i^2 + \frac{1}{2}mu_i^2\}\,\Delta t \\ &= \mathcal{J}\underbrace{(u_0, u_1, \cdots, u_{n-1})}_{=:\underline{u}} = \mathcal{J}(\Big\{u_k\Big\}_{k=0}^{n-1}) \end{aligned} \tag{14}$$

hence

$$\mathcal{J}(\underline{u}) = \sum_{i=1}^{n}\{\frac{1}{2}I_i^2(\underline{u}) + \frac{1}{2}mu_i^2\}\,\Delta t \tag{15}$$

so the optimization problem reduces to optimizing $\mathcal{J}$ in respect to $\underline{u}$, i.e.

$$\frac{\partial \mathcal{J}(\underline{u})}{\partial u_j} = 0 \tag{16}$$

for all u_j.

After some calculations one finds the following expressions for the partial derivatives fo the cost function $\mathcal{J}$

$$\frac{\partial \mathcal{J}(\underline{u})}{\partial u_j} = mu_j \cdot \Delta t + \sum_{i=j+1}^{n+1} I_i(\underline{u}) \cdot (-(N - I_0)B_i)\,(\Delta t)^2 \tag{17}$$

with $I_i(\underline{u}) = N - (N - I_0)B_i$ and the B_i being functions of the control signal $\underline{u}$. These partial derivatives either could be analyzed to be zero, but no solution is easily visible, or they can be used for a gradient method numerically, the way we went finally. Hence for the update of the $(\nu + 1)$ step of the control signal u_j from the ν step with initially zero control $u_j^{(0)} := 0$ is given by

$$u_j^{(\nu+1)} = u_j^{(\nu)} + h_j \tag{18}$$

with $h_j = c \cdot (-\partial\mathcal{J}/\partial u_j)$, i.e. maximizing $G := -\mathcal{J}$, and a step size c initially quite large, and then halfing every time it overshoots the minimization process. Numerical results will be presented at the CMMSE 2016 conference.

2.3 Generalization to not directly solvable disease dynamics

In cases when the disease dynamics $\dot{I} = f(I, u)$ cannot be solved analytically, we can use Lagrange multipliers $\lambda(t)$ for the optimization of the cost function, taking the disease dynamics as constraint into account, hence

$$\mathcal{L}[I(t), u(t), \lambda(t)] = \int_{t_0}^{T} \{\frac{1}{2}I^2 + \frac{1}{2}mu^2 + \lambda(\dot{I} - f(I, u))\}\, dt \tag{19}$$

to be minimized.

Now the discretization, done in the same way as before, reveals some interesting insights, namely the variation in respect to the Lagrange multipiers $\partial\mathcal{L}/\partial\lambda_j = 0$ gives back the constraint of the disease dynamcis

$$\frac{1}{\Delta t}(I_{j+1} - I_j) = f(I_j, u_j) \tag{20}$$

and then variation in respect to the infecteds $\partial\mathcal{L}/\partial I_j = 0$ gives a backward in time dynamics for the Lagrange multipliers

$$\frac{1}{\Delta t}(\lambda_{j-1} - \lambda_j) = \lambda_j \frac{\partial f}{\partial I_j} - I_j \tag{21}$$

with the upper boundary condition $\lambda_n = -I_{n+1} \cdot \Delta t \to 0$ for small time steps $\Delta t \to 0$. Finally the variation in respect to the control signal $\partial\mathcal{L}/\partial u_j = 0$ gives an algebraic equation system

$$mu_j - \lambda_j \frac{\partial f}{\partial u_j} = 0 \tag{22}$$

which in the case of a linear control in f, like we have here in the linear infection model, gives the control signal u_j as a function of λ_j and I_j only, due to $\partial f/\partial u_j$ is then only a function of I_j, and not of u_j any more. Hence then the forward dynamics for the infected and the backward dynamics for the Lagrange multipliers are sufficient to solve the optimal control problem. Again, numerically this can be done by a gradient method as described above.

2.4 Shooting method

Though the gradient method can be applied to the forward/backward dynamics of disease dynamics and Lagrange multipliers, it is somehow inelegant, sticking with a backward iteration. This drawback can be overcome by a shooting method, where also the Lagrange multiplier dynamics is iterated forward, starting from an arbitrary (but numerically eventually already well suited) initial condition λ_0, arriving at an end point $\lambda_n \neq 0$, but then

via Newton's method iteratively seek a solution λ_0 for which the function $\lambda_n(\lambda_0)$ vanishes, hence $\lambda_n(\lambda_0) = 0$, hence

$$\lambda_0^{(\nu+1)} = \lambda_0^{(\nu)} - \frac{\lambda_n(\lambda_0^{(\nu)})}{\frac{\partial \lambda_n}{\partial \lambda_0}} \tag{23}$$

to be used numerically from step ν to step $(\nu + 1)$.

Acknowledgements

This work has been supported by the European Union under FP7 in the DENFREE project, and the kind support by the Wellcome Trust and by FCT, Portugal, in various ways. We also thank the DAAD and FCT to support the interaction between University of Koblenz and University of Lisbon in an DAAD-FCT exchange grant.

References

[1] D.Aldila, E.Soewono, N.Nuraini, *On the Analysis of Effectivness in Mass Application of Mosquito Repellent for Dengue Disease Prevention*, *AIP Conf. Proc.* **1450** (2012), 103–109

[2] Karunia Putra Wijaya, Thomas Goetz, Edy Soewono, *An Optimal Control Model of Mosquito Reduction Managment in Dengue Endemic Region*, *International Journal of Biomathematics* **Vol. 7, No. 5** (2014) 1450056, 22 pages, *DOI: 10.1142/SI793524514500569*

[3] Luís Mateus, Maíra Aguiar and Nico Stollenwerk (2015) Bayesian estimation of vaccine efficacy. *Proceedings of the 15th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2015, Cadiz, Spain*, pp. 794–802, ISBN: 978-84-617-2230-3, edited by Jesus Vigo et al.

[4] M. R. Capeding et al., *Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy childen in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial*, Lancet **384** (2014) 1358–65.

[5] Rocha, F., Aguiar, M., Souza, M., & Stollenwerk, N. (2013) Time-scale separation and center manifold analysis describing vector-borne disease dynamics, *Int. Journal. Computer Math.* **90**, 2105–2125.

[6] Stollenwerk, N., & Jansen, V. (2011) *Population Biology and Criticality: From critical birth–death processes to self-organized criticality in mutation pathogen systems* (Imperial College Press, World Scientific, London).

[7] Aguiar, M., Stollenwerk, N., & Halstead, S. (2016) The impact of the newly licensed dengue vaccine in endemic countries, *submitted for publication.*

[8] Aguiar, M., Halstead, S., & Stollenwerk, N. (2016) DengVaxia: improving disease burden reduction, *submitted for publication.*

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Paramagnetic H-related defects in silica: a first-principles investigation.

Luigi Giacomazzi[1], L. Martin-Samos[2] and N. Richard[3]

[1] *CNR-IOM/Democritos, via Bonomea 265, 34136 Trieste (Italy)*

[2] *MRL, University of Nova Gorica, Vipavska 11c 5270-Ajdovscina, (Slovenija)*

[3] *CEA, DAM, DIF, F-91297 Arpajon (France)*

emails: giacomaz@sissa.it, lmartinsamos@gmail.com, nicolas.richard@cea.fr

Abstract

Due to its ubiquitous presence, high diffusivity, and reaction capabilities, hydrogen represents an important source of absorbing centers in optical fibers. Despite its well known darkening effect, the structural details and generation mechanisms of some H-related defect as e.g. the E'_β are not yet well understood. In this paper we apply first-principle state-of-the-art techniques to investigate issues concerning H-related paramagnetic centers in silica.

Key words: E'_β, silica, EPR

1 Introduction

The use of optical fibers has revolutionized the telecommunications industry, in particular, by allowing an enhancement of the data transmission speed of four order of magnitude in the last 30 years [1]. Because of the rapidly increasing traffic in core networks future optical fibers are expected to require a transmission capacity well beyond the current optical fiber input power limit [1]. In such a context, reducing the presence of absorbing centers (i.e. defects that can be intrinsic as e.g. oxygen vacancies or extrinsic as e.g. hydrogen impurity) will become a crucial task for the optimization of the future optical fibers. Hydrogen darkening is a rather common phenomenon of signal attenuation. The latter can be caused either by molecular hydrogen or defect sites that have reacted with hydrogen to form other absorbing centers. Typical hydrogen sources include not only the ambient atmospheric

 ISBN: 978-84-608-6082-2

hydrogen but also the one coming from galvanic reaction between dissimilar metals in the cable or even from the action of certain microorganisms (e.g. sulfate-reducing bacteria).

In his review of defects in silica, Griscom [2] discussed the main paramagnetic defects related to the hydrogen presence observed using γ- or 100 KeV X-rays. The first one is the radiolytic atomic hydrogen (H^0) which exhibits a Fermi contact of 50.5 mT and a almost isotropic g tensor ($g \sim 2.0025$). The radiolytic hydrogen by reacting with a proper precursor site (i.e. an oxygen vacancy) should transform in to the so-called E'_β center [2]:

$$H^0 + [\equiv Si - Si \equiv] \longrightarrow [\equiv Si - H + \equiv Si\cdot] = E'_\beta \tag{1}$$

where $\cdot$ indicated an unpaired electron and $\equiv$Si indicates a three-fold Si atom. As the EPR spectrum of the E'_β does not show hyperfine interaction with a proton it follows that a separation of at least 5 Å should occur between the proton "H" and the spin "$\cdot$" [2]. Furhermore since the oxygen vacancy could occur as a twofold Si atom one needs also to consider the following reaction (Eq. 2): $H^0 + [=Si:] \longrightarrow [=Si - H]\cdot$
which is known as the H(I) center [2]. While the structural model of the latter H(I) center has received several confirmations both from theory and experiments the former model [Eq. (1)] of the E'_β center has received very little attention and, to our knowledge, no theoretical attempt has been done to discuss such a model. In the present paper we discuss the structural origin of the above mentioned types of H-related centers in silica glass. To this aim we show and discuss accurate state-of-the-art first-principles calculations for a significant set of models that includes the most considered representatives for the observed paramagnetic H-related centers, i.e. H^0, the E'_β as in Eq. (1), and the H(I) centers as in Eq. (2).

2 Metods and models details

The EPR calculations presented in this work are based on Density Functional Theory (DFT). In particular we adopt the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional. We use norm-conserving, scalar-relativistic Trouiller-Martins gipaw pseudopotentials and Kohn-Sham wavefunctions are expanded in a basis of plane waves up to a kinetic cutoff of 80 Ry [3]. The adopted amorphous silica model consists in a supercell containing 108 atoms at a density close to the experimental one ~ 2.2 g/cm^3 [5]. We obtain hydrogen-doped silica configurations by inserting a H atom in a neutrally charged oxygen vacancy configuration that was generated starting from a pure a-SiO_2 model [5]. We similarly generated other two configurations (neutrally charged) by adding, in two different places, a H atom to the unpuckered configuration that was obtained for the same vacancy site (Fig. 1). Moreover by inserting a H_2 molecule (H-H distance 0.75 Å) inside the positively charged oxygen vacancy we could obtain a configuration where the vacancy is passivated by a H atom while the other one becomes an interstitial H atom (H^0), as also confirmed by our EPR

 ISBN: 978-84-608-6082-2

parameters calculations. The configurations of the H(I) centers (Eq. 2) were generated by adding a H atom placed near a twofold Si atom [6]. All the configurations mentioned here above have been relaxed by first-principles prior to the calculation of the EPR parameters. We use the QE-GIPAW code, which implements the gauge including projector augmented wave (GIPAW) method [3], of the Quantum ESPRESSO package [4] for the calculation of the EPR parameters (EPR hyperfine and g-tensors) of the here investigated defect models.

3 Results

In Fig. 1(a) we show the configuration obtained by following directly the left hand side of Eq. 1. During the relaxation the two ≡Si units rotate to accomodate the hydrogen atom but no major structural change occurs and the spin-density points towards the center of the oxygen vacancy. The hydrogen atom shows some spin localization which is confirmed also by the non-negligible Fermi contact A_{iso}(H)= 1.3 mT. As the latter was not found in the experiments [2] we should rule out Fig. 1(a) as a possible candidate of the E'_β. Next, by taking advantage of a previously generated unpuckered configuration [5], we tried to directly obtain a model as suggested by the right hand side of Eq. 1 and in particular we started first by generating a configuration [Fig. 1(b)] which resembles very much the E'_β model given in Fig. 3 of [2]. The Fermi contact $A_{iso}(^{29}Si)$ is almost 2 mT lower than found in the H-free model [5]. The g-tensor is slightly more axial than in the H-free model as suggested by the small decrease in the $g_{23} = g_2 - g_3$ value. However the g_2 value is slightly larger than found for the H-free model in contrast to the decrease observed between the g_2 values of the E'_β and E'_γ centers [2]. Because of the above listed discrepancies between our calculations, based on the model given in Fig. 3 of [2], and the experimental data we have calculated the EPR parameters of the configuration where the hydrogen and the unpaired electron exchange their place [Fig. 1(c)]. Such a configuration shows a $A_{iso}(^{29}Si)$ larger by 5.5 mT with respect to the one of the H-free model and also a slightly decreased g_2 value, even if the g_{23} value increases with respect to Fig. 1(b). Furthermore the average of the EPR parameters obtained for Fig. 1(b) and (c) suggests that a better representation of the E'_β could be achieved by considering not only Fig. 1(b) but also other H-doped configurations like e.g. Fig. 1(c).

The H(I) center is experimentally characterized by A_{iso}(H)∼7.4 mT and g tensor $g_1 = 2.0022$, $g_2 = 2.0016$, $g_3 = 2.0003$ [2]. On average our calculations provide Fermi contacts and g-values differing with experiments by less than 0.4 mT and $3{\cdot}10^{-4}$ respectively.

4 Conclusions

We performed a preliminary series of first-principles calculations of the EPR parameters for a set of hydrogen-doped amorphous silica models. In particular, the present calculations

 ISBN: 978-84-608-6082-2

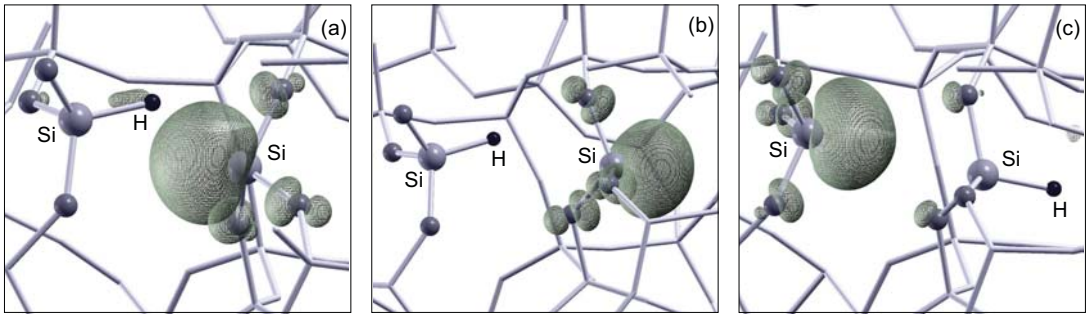


Figure 1: Spin densities (shadowed grey) obtained by a first-principles relaxation of neutrally charged H-doped silica configurations: (a) the H atom is first placed at about the middle position of the Si-Si, (b) and (c) the H atom is placed near one or the other of the two Si atoms of an unpuckered configuration [5]. H, O and Si atoms are shown in black, dark grey and grey respectively.

suggest that the commonly accepted structural model of the E'_β should be updated by considering at least two alternative sites hosting the hydrogen atom.

References

[1] T. Morioka, M. Jinno, H. Takara, and H. Kubota, *Innovative Future Optical Transport Network Technologies*, NTT Technical Review **9** (2011) No. 8.

[2] D. L. Griscom, *Optical properties and structure of defects in silica glass*, J. Ceram. Soc. Japan **99** (1991) 923-942.

[3] C.J. Pickard and F. Mauri, *First-principles theory of the EPR g tensor in solids: Defects in quartz*, Phys. Rev. Lett. **88** (2002) 086403.

[4] P. Giannozzi *et al.*, *QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials*, J. Phys.: Condens. Matter **21** (2009) 395502.

[5] L. Giacomazzi, L. Martin-Samos, A. Boukenter, Y. Ouerdane, S. Girard, N. Richard, *EPR parameters of E' centers in v-SiO_2 from first-principles calculations*, Phys. Rev. B **90** (2014) 014108.

[6] L. Giacomazzi, L. Martin-Samos, A. Boukenter, Y. Ouerdane, S. Girard, N. Richard, *Ge(2), Ge(1) and Ge-E' centers in irradiated Ge-doped silica: a first-principles EPR study*, Opt. Mater. Express **5** (2015) 1054-1064.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Two-body interactions and the physics of natural occupation numbers and amplitudes

Klaas J. H. Giesbertz[1] **and Robert van Leeuwen**[2]

[1] *Theoretical Chemistry, Faculty of Exact Sciences, VU University, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands*

[2] *Department of Physics, Nanoscience Center, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Survontie 9, Jyväskylä, Finland*

emails: k.j.h.giesbertz@vu.nl, robert.vanleeuwen@jyu.fi

Abstract

The non-vanishing of the natural orbital occupation numbers of the one-particle density matrix of many-body systems has important consequences for the existence of a density matrix-potential mapping for nonlocal potentials in reduced density matrix functional theory and for the validity of the extended Koopmans' Theorem. We analyse for a number of explicit examples of two-particle systems that in case the wave function is non-analytic at its spatial diagonal and derive a more general criterium for the non-vanishing of natural occupations for two-particle wave functions with a certain separability structure. Singlet two-electron systems also allow for a diagonal representation of the wave function with the natural orbitals as its eigenfunctions and the squares of its eigenvalues, the natural amplitudes, are equal to the occupation numbers. We demonstrate that the sign pattern of the natural amplitudes is related to the long-range structure of the wave function, which is in turn dictated by the tail of the Coulomb interaction.

Key words: one-body reduced density matrix, vanishing, natural occupation number, natural amplitude, decay

The one-body density matrix of a wavefunction Ψ can be defined in terms of the usual field operators as

$$\gamma(\mathbf{x}, \mathbf{x}') \coloneqq \langle \Psi | \hat{\psi}^\dagger(\mathbf{x}') \hat{\psi}(\mathbf{x}) | \psi \rangle \,, \tag{1}$$

  ISBN: 978-84-608-6082-2

where $\mathbf{x} \coloneqq \mathbf{r}\sigma$ is space-spin coordinate. The natural occupation numbers are defined as the eigenvalues of the one-body reduced density matrix (1RDM)

$$\int \mathrm{d}\mathbf{x}' \, \gamma(\mathbf{x}, \mathbf{x}')\phi_k(\mathbf{x}') = n_k \, \phi_k(\mathbf{x}) \, . \tag{2}$$

The corresponding eigenfunctions are called the natural orbitals (NOs) [1].

The sum of the occupation numbers equals the number of electrons in the system. Therefore, if we order the occupation numbers, n_k, from the highest to the lowest one, their values need to decay to zero sufficiently fast for $k \to \infty$, i.e.

$$\lim_{k\to\infty} n_k = 0 \, , \tag{3}$$

or even become zero after some point $k_{\max}$. The question whether they actually do become zero or only approach zero for $k \to \infty$ is not only an academic question, but also of practical interest for methods that try to build an accurate approximation to the wavefunction by making an expansion in terms of Slater determinants, e.g. configuration interaction (CI). The non-vanishing of the natural orbital (NO) occupation numbers of the one-particle density matrix of many-body systems also has important consequences for the existence of a density matrix-potential mapping for nonlocal potentials in reduced density matrix functional theory [2, 3] and for the validity of the extended Koopmans' theorem [4, 5, 6, 7].

To investigate this matter [8, 9, 10], we limit ourselves to singlet two-electron system. Since the spatial part of the wavefunction is symmetric, it can be diagonalised

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \sum_k c_k \, \phi_k(\mathbf{r}_1)\phi_k(\mathbf{r}_2) \, . \tag{4}$$

By calculating the corresponding spin-integrated 1RDM, one readily finds that the eigenfunctions are NOs. The eigenvalues c_k are called the natural amplitudes and are related to the occupation numbers as $n_k = c_k^2$. The advantage is that we can access the NOs and natural occupations directly from the wavefunction, bypassing the construction of the 1RDM (1).

Since the wavefunction has a cusp at the coalescence points of the two electrons, the first order derivative of the wavefunction is discontinuous. Combined with a theorem by Weyl [11], this allows us to put a modest lower bound to the decay rate of the natural amplitudes

$$\lim_{k\to\infty} |c_k| k^{1/2} = 0 \, . \tag{5}$$

On the other hand, an infinitely differentiable wavefunction can be shown to have a much higher decay rate of its natural amplitudes with the help of a result by Hille and Tamarkin [12, 13]

$$\lim_{k\to\infty} |c_k| R^{k/4} = 0 \, , \tag{6}$$

 ISBN: 978-84-608-6082-2

where R is some constant.

Equations (5) and (6) only provide lower bounds to the decay rate and do not say much on the more difficult question whether the occupation numbers do become zero or not. For a two-particle wavefunction of the special form

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \alpha(\mathbf{r}_1)\alpha(\mathbf{r}_2)f(\mathbf{r}_1 - \mathbf{r}_2)\,, \tag{7}$$

we have been able to establish the following theorem.

Theorem 1. *A two-particle wavefunction of the form* (7) *has no vanishing natural amplitudes (eigenvalues) if and only if the Fourier transform of f is nonzero almost everywhere.*

A detailed proof is exhibited in Ref. [8]. The use of this theorem is obvious, since one does not need to explicitly construct the NOs and natural amplitudes for wavefunctions of the form (7) to determine whether any of the natural occupation numbers are zero. For example, we can show that for a two-particle system of harmonically confined electrons with a Coulombic interaction, the so-called Hookium [14, 15], the natural occupation numbers never vanish. We have also used this proof to show that a fully optimised wavefunction of the form (7) for the 1D H_2 molecule with soft Coulomb interactions does not have vanishing occupation numbers [10].

The natural amplitudes do not only determine the natural occupation numbers, but also cary a phase, which can only be $+1$ or -1 for a real singlet ground state wavefunction. In the second part we explore the relation between the signs of the natural amplitudes and the nature to the two-body interaction [9]. We show that long-range Coulomb- type interactions are responsible for the appearance of multiple positive amplitudes.

We also explore how the sign pattern changes when the system parameters are changed, e.g. the bond length in the H_2 molecule or the confinement of the Hookium. It has been argued by Goedecker and Umrigar that in weakly correlated systems, e.g. the He atom and Hookium with strong confinement, that if the phase of highest occupied NO is positive, all the other natural amplitudes will be negative [16]. In the strongly correlated regime, however, the sign pattern becomes alternating, as demonstrated explicitly for the Hookium [17] and the hydrogen molecule [18]. It therefore appears that when the bond in the H_2 molecule is stretched or the parabola in the Hookium is widened, that the natural amplitudes need to cross zero.

We will argue that this is actually not the case. Firstly, the argument by Goedecker and Umrigar is based on the assumption that the all the NOs resemble the Hartree–Fock orbitals. This assumption is obviously wrong. That the statement of Goedecker and Umrigar is incorrect, has indeed been demonstrated by very large basis set calculations on the He atom for which multiple positive natural amplitudes were found [19]. Nevertheless, the sign pattern changes from predominantly negative to alternating when changing from the weakly to the strongly correlated regime, so one could surmise that the natural amplitudes

need to cross zero to change their sign. However, this argument is only valid in a finite basis. In an infinite basis, an infinite amount of natural amplitudes will be clustered around zero. Since eigenvalues typically can only cross under very special circumstances, natural amplitudes can typically never become zero. The simple reason is that there always will be a natural amplitude closer to zero, blocking its approach. This mechanism of avoided crossings therefore prevents the natural amplitudes to become zero when variations in the external potential are made, e.g. nuclear configuration of a molecule.

The only way natural amplitudes can cross zero is when all natural amplitudes go together through zero. A combined collapse only occurs by making modifications to the interaction itself. For the Coulomb interaction this be be the non-interacting limit. For special interactions such a collapse of the natural amplitudes can also occur at finite interaction strength. An example is the Hamiltonian of the form

$$\hat{H} = -\frac{1}{2}\nabla^2_{\mathbf{r}_1} - \frac{1}{2}\nabla^2_{\mathbf{r}_2} + \frac{1}{2}\omega^2(r_1^2 + r_2^2) + \frac{\lambda}{r_{12}^2}. \tag{8}$$

The ground state of this Hamiltonian has the following surprisingly simple form [20]

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = N_\alpha \, \mathrm{e}^{-\frac{1}{2}\omega(r_1^2+r_2^2)} r_{12}^\alpha, \tag{9}$$

where $\alpha = \left(\sqrt{1+4\lambda} - 1\right)/2$ and N_α is a normalisation factor. Exactly when α is an even integer, most of the natural amplitudes collapse to zero and only a finite amount remain nonzero [9].

The same line of reasoning immediately carries over to the natural occupation numbers, so we have a very strong argument that the occupation numbers in systems with a Coulomb interaction never vanish. Note that this argument only applies to an infinite Hilbert space, so is not applicable to the model Hamiltonians due to the finite basis representation on our computers. In a finite Hilbert space there will be a smallest occupation number which can vanish, since there is no lower occupied NO anymore that can prevent this, via an avoided crossing.

Acknowledgements

The authors acknowledge the Academy of Finland for research funding under Grant No. 127739. KJHG also gratefully acknowledges a VENI grant by the Netherlands Foundation for Research NWO (722.012.013).

References

[1] P.O. Löwdin, Phys. Rev. **97**, 1474 (1955).

[2] T.L. Gilbert, Phys. Rev. B **12**, 2111 (1975).

[3] T. Baldsiefen, A. Cangi, and E.K.U. Gross, Phys. Rev. A **92**, 052514 (2015).

[4] D.W. Smith and O.W. Day, J. Chem. Phys. **62**, 113 (1975).

[5] O.W. Day, D.W. Smith, and R.C. Morrison, J. Chem. Phys. **62**, 115 (1975).

[6] M.M. Morrell, R.G. Parr, and M. Levy, J. Chem. Phys. **62**, 549 (1975).

[7] J. Katriel and E.R. Davidson, Proc. Natl. Acad. Sci. USA **77**, 4403 (1980).

[8] K.J.H. Giesbertz and R. van Leeuwen, J. Chem. Phys. **139**, 104109 (2013).

[9] K.J.H. Giesbertz and R. van Leeuwen, J. Chem. Phys. **139**, 104110 (2013).

[10] K.J.H. Giesbertz and R. van Leeuwen, J. Chem. Phys. **140**, 184108 (2014).

[11] H. Weyl, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen **2**, 110 (1911).

[12] E. Hille and J.D. Tamarkin, Acta Math. **57**, 1 (1931).

[13] J.M. Rasmussen, *Compact Linear Operators and Krylov Subspace Methods*, Master's thesis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark (2001).

[14] N.R. Kestner and O. Sinanoḡlu, Phys. Rev. **128**, 2687 (1962).

[15] M. Taut, Phys. Rev. A **48**, 3561 (1993).

[16] S. Goedecker and C.J. Umrigar, *Many-Electron Densities and Reduced Density Matrices*, Chap. 8, pp. 165–181 (Kluwer Academic, Dordrecht - New York, 2000).

[17] J. Cioslowski and K. Pernal, J. Chem. Phys. **113**, 8434 (2000).

[18] J. Cioslowski and K. Pernal, Chem. Phys. Lett. **430**, 188 (2006).

[19] X.W. Sheng, Ł.M. Mentel, O.V. Gritsenko, and E.J. Baerends, J. Chem. Phys. **138**, 164105 (2013).

[20] L.D. Landau and E.M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory*, Vol. 3 of *Course of Theoretical Physcis* (Pergamon Press, Oxford, 1977), 3rd edition.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Exact solutions of the Schrödinger equation for two electrons on a sphere

Peter M.W. Gill[1] **and Pierre-François Loos**[1]

[1] *Research School of Chemistry, Australian National University*

emails: peter.gill@anu.edu.au, pf.loos@anu.edu.au

Abstract

We show that simple, closed-form wavefunctions and energies that can be found for two Coulombically interacting electrons confined to a D-dimensional sphere. After outlining the method for solving the relevant Schrödinger equations, we give particular solutions for the cases of a ring ($D = 1$), a normal sphere ($D = 2$) and a glome ($D = 3$).

Key words: Exactly soluble systems, two-electron systems, ringium, spherium, glomium

1 Introduction

It is often said that it is impossible to find exact wavefunctions and energies for a system with two or more electrons. This is because, in general, the associated Schrödinger equation

$$\left[-\frac{\nabla_1^2}{2} - \frac{\nabla_2^2}{2} + V(\mathbf{r}_1) + V(\mathbf{r}_2) + \frac{1}{r_{12}}\right]\Psi = E\Psi \tag{1}$$

(where V is the external potential and $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the interelectronic distance) is not separable in any convenient coordinate system. Fortunately, however, this is not always true and there exist a few special systems where separation is possible and surprisingly simple and elegant solutions can be obtained. Many years ago, Kais et al. showed how to achieve this for a pair of electrons in a harmonic well [1] and Taut subsequently generalized this to a countably infinite number of confinement strengths [2]. Taut later extended his approach to cases wherein an external magnetic field is also present [3]. These papers have been well cited, largely because exact solutions are ideal testbeds [4, 5, 6, 7, 8, 9, 10, 11] for the refinement of cruder models like density functional theory (DFT) and quantum Monte Carlo (QMC).

 ISBN: 978-84-608-6082-2

A few years ago, we discovered that the Schrödinger equation for two electrons confined to a D-dimensional sphere (a ring ($D = 1$), a normal sphere ($D = 2$), a glome ($D = 3$), etc.) can also be solved in closed form for certain special values of the sphere radius R.

2 Two electrons on a D-sphere

2.1 Two electrons on a ring

In the $D = 1$ case, the Hamiltonian is

$$\hat{H} = -\frac{1}{2R^2}\left[\frac{\partial^2}{\partial\theta_1^2} + \frac{\partial^2}{\partial\theta_2^2}\right] + \frac{1}{R\sqrt{2-2\cos(\theta_1-\theta_2)}} \tag{2}$$

In the coordinates $\Omega = (\theta_1+\theta_2)/2$ and $u = R\sqrt{2-2\cos(\theta_1-\theta_2)}$, the Hamiltonian separates into the sum of

$$\hat{H}_\Omega = -\frac{1}{4R^2}\frac{d^2}{d\Omega^2} \qquad\qquad \hat{H}_u = \left[\frac{u^2}{4R^2} - 1\right]\frac{d^2}{du^2} + \frac{u}{4R^2}\frac{d}{du} + \frac{1}{u} \tag{3}$$

and the total wavefunction and total energy are then given by

$$\Phi(\Omega, u) = \Lambda(\Omega)\Psi(u) \qquad\qquad E = \mathcal{E} + \epsilon \tag{4}$$

The eigenfunctions and eigenvalues of $\hat{H}_\Omega$ are

$$\Lambda_J(\Omega) = \exp(iJ\Omega) \qquad\qquad \mathcal{E}_J = \frac{J^2}{4R^2} \tag{5}$$

The eigenfunctions of H_u satisfy a Heun-like equation and fall into four families [12]. For a countably infinite number of special values of the radius R, the eigenfunctions reduce to elementary functions. For example, for the indicated radii, one finds the ground states

$$R = 1/2 \qquad \epsilon = 9/4 \qquad \Psi(u) = u\sqrt{1+u} \tag{6}$$

$$R = \sqrt{6}/2 \qquad \epsilon = 2/3 \qquad \Psi(u) = u(2+u) \tag{7}$$

$$R = \sqrt{46}/2 \qquad \epsilon = 9/46 \qquad \Psi(u) = u(92+46u+5u^2) \tag{8}$$

and the first excited states

$$R = \sqrt{10}/2 \qquad \epsilon = 9/10 \qquad \Psi(u) = u(2+u)\sqrt{10-u^2} \tag{9}$$

$$R = \sqrt{66}/2 \qquad \epsilon = 8/33 \qquad \Psi(u) = u(132+66u+7u^2)\sqrt{66-u^2} \tag{10}$$

A method for obtaining these elementary solutions, along with several additional examples, can be found elsewhere [12].

 ISBN: 978-84-608-6082-2

2.2 Two electrons on a sphere

In the $D = 2$ case, the Hamiltonian again separates after a change of coordinates. By following Breit's approach [13], the spin and angular wavefunctions can be factorized out, leaving the univariate interelectronic Hamiltonian $\hat{H}_u$.

Within the manifold of ${}^1S$ states, the interelectronic Hamiltonian is

$$\hat{H}_u = \left[\frac{u^2}{4R^2} - 1\right] \frac{d^2}{du^2} + \left[\frac{3u}{4R^2} - \frac{1}{u}\right] \frac{d}{du} + \frac{1}{u} \tag{11}$$

and, as on a ring, there are an infinite number of special values of R for which the eigenfunctions are elementary. For example, for the indicated radii, one finds the ground states

$$R = \sqrt{3}/2 \qquad \epsilon = 1 \qquad \Psi(u) = 1 + u \tag{12}$$

$$R = \sqrt{7} \qquad \epsilon = 2/7 \qquad \Psi(u) = 28 + 28u + 5u^2 \tag{13}$$

Within the manifold of ${}^3P$ states, the interelectronic Hamiltonian is

$$\hat{H}_u = \left[\frac{u^2}{4R^2} - 1\right] \frac{d^2}{du^2} + \left[\frac{5u}{4R^2} - \frac{3}{u}\right] \frac{d}{du} + \frac{1}{u} \tag{14}$$

and some of the ground-state solutions include

$$R = \sqrt{15}/2 \qquad \epsilon = 1/3 \qquad \Psi(u) = 3 + u \tag{15}$$

$$R = \sqrt{23} \qquad \epsilon = 3/23 \qquad \Psi(u) = 276 + 92u + 7u^2 \tag{16}$$

A method for obtaining these elementary solutions, along with several additional examples, can be found elsewhere [14, 15].

2.3 Two electrons on a glome

In the $D = 3$ case, we follow the same approach as for the $D = 2$ case.

Within the ${}^1S$ manifold, the interelectronic Hamiltonian is

$$\hat{H}_u = \left[\frac{u^2}{4R^2} - 1\right] \frac{d^2}{du^2} + \left[\frac{5u}{4R^2} - \frac{2}{u}\right] \frac{d}{du} + \frac{1}{u} \tag{17}$$

and examples of ground-state solutions include

$$R = \sqrt{5/2} \qquad \epsilon = 1/2 \qquad \Psi(u) = 2 + u \tag{18}$$

$$R = \sqrt{33/2} \qquad \epsilon = 2/11 \qquad \Psi(u) = 132 + 66u + 7u^2 \tag{19}$$

Within the ${}^3P$ manifold, the interelectronic Hamiltonian is

$$\hat{H}_u = \left[\frac{u^2}{4R^2} - 1\right] \frac{d^2}{du^2} + \left[\frac{7u}{4R^2} - \frac{4}{u}\right] \frac{d}{du} + \frac{1}{u} \tag{20}$$

 ISBN: 978-84-608-6082-2

and examples of ground-state solutions include

$$R = \sqrt{14/2} \qquad \epsilon = 1/4 \qquad \Psi(u) = 4 + u \tag{21}$$

$$R = \sqrt{77/2} \qquad \epsilon = 8/77 \qquad \Psi(u) = 616 + 154u + 9u^2 \tag{22}$$

A method for obtaining these elementary solutions, along with several additional examples, can be found elsewhere [14].

Acknowledgements

This work was supported by grants (DP0664466, DP0771978, DP0984806, DP1094170, DP120104740, DE130101441 and DP140104071) from the Australian Research Council.

References

[1] S. Kais, D. R. Herschbach and R. D. Levine, *Dimensional scaling as a symmetry operation*, J. Chem. Phys. **91** (1989) 7791–7796.

[2] M. Taut, *Two electrons in an external oscillator potential: Particular analytic solutions of a Coulomb correlation problem*, Phys. Rev. A **48** (1993) 3561–3566.

[3] M. Taut, *Two electrons in a homogeneous magnetic field: particular analytical solutions*, J. Phys. A: Math. Gen. **27** (1994) 1045–1055.

[4] C. Filippi, C. J. Umrigar and M. Taut, *Comparison of exact and approximate density functionals for an exactly soluble model*, J. Chem. Phys. **100** (1994) 1290–1296.

[5] C.-J. Huang and C. J. Umrigar, *Local correlation energies of two-electron atoms and model systems*, Phys. Rev. A **56** (1997) 290–296.

[6] Z. X. Qian and V. Sahni, *Physics of transformation from Schrodinger theory to Kohn-Sham density-functional theory: Application to an exactly solvable model*, Phys. Rev. A **57** (1998) 2527–2538.

[7] M. Taut, A. Ernst and H. Eschrig, *Two electrons in an external oscillator potential: exact solution versus one-particle approximations*, J. Phys. A: At. Mol. Opt. Phys. **31** (1998) 2689–2708.

[8] J. Cioslowski and K. Pernal, *The ground state of harmonium*, J. Chem. Phys. **113** (2000) 8434–8443.

[9] J. H. Lloyd-Williams, R. J. Needs and G. J. Conduit, *Pseudopotential for the electron-electron interaction*, Phys. Rev. B **92** (2015) 075106/1–8.

[10] P. F. Loos, N. J. Bloomfield and P. M. W. Gill, *Three-electron coalescence points in two and three dimensions*, J. Chem. Phys. **143** (2015) 181101/1–4.

[11] J. Cioslowski, M. Piris and E. Matito, *Robust validation of approximate 1-matrix functionals with few-electron harmonium atoms*, J. Chem. Phys. **143** (2015) 214101/1–10.

[12] P. F. Loos and P. M. W. Gill, *Exact wave functions of two-electron quantum rings*, Phys. Rev. Lett. **108** (2012) 083002/1–4.

[13] G. Breit, *Separation of angles in the two-electron problem*, Phys. Rev. **35** (1930) 569–578.

[14] P. F. Loos and P. M. W. Gill, *Two electrons on a hypersphere: A quasi-exactly solvable model*, Phys. Rev. Lett. **103** (2009) 123008/1–4.

[15] P. F. Loos and P. M. W. Gill, *Excited states of spherium*, Mol. Phys. **108** (2010) 2527–2532.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A Hybrid GPU Technique for Real-Time Terrain Visualization

Cesar González[1], Mariano Pérez[1] and Juan M. Orduña[1]

[1] *Departmento de Informática, Universidad de Valencia*

emails: cegonse@alumni.uv.es, mariano.perez@uv.es, juan.orduna@uv.es

Abstract

Real-Time terrain visualization plays an important rule in multiple popular applications like geographical information systems, computer games, or civil or militar simulators, where hardware tessellation has become a de-facto standard nowadays in the graphic pipeline. Also, post-processing techniques enhance the appearance of the rendered image by applying changes at the pixel level using the fragment shader, without increasing the number of polygons, but they have not been still used in terrain rendering due to different reasons. In this paper, we present a new real-time terrain rendering approach which efficiently combines hardware tessellation and parallax mapping, making parallax mapping compatible with hardware tessellation and terrain rendering. The performance evaluation results show that the proposed scheme improves the performance of real-time terrain rendering applications in regard to the performance yielded when exclusively using hardware tessellation.

Key words: Terrain visualization, Real-Time rendering, GPU shaders, Hardware tessellation, Parallax mapping)

1 Introduction

Real-Time terrain visualization is a very active research field in the area of computer graphics and plays an important rule in multiple applications like Geographic Information Systems (GIS) [13, 14], computer games [2] or civil or militar simulators [1]. Figure 1 shows an example of an image displayed by a GIS application, a mountain valley with some villages.

One of the main tasks of these applications is to display a high visual quality terrain model at interactive frame rates. Terrain datasets used in these applications usually exceed the rendering capabilities of currently available hardware graphics, and their interactive

Figure 1: Example image displayed by a Geographic Information System application

rendering require that applications adjust their complexity in a view-dependent manner. Traditionally, most of the techniques managed the level-of-detail required at every point of the terrain surface by executing CPU-based algorithms [4]. However, modern GPU features provide a more efficient way to control view-dependent level-of-detail using GPU-based algorithms [5, 6, 7, 8, 9].

One of the features that are potentially useful for level-of-detail control appeared in the Shader Model version 4: the geometry shader stage[10]. This new stage in the graphics pipeline allows the creation of view-dependent geometry directly on the GPU without transferring all the data from the CPU, thus increasing the rendering frame-rate and visual quality. However, this feature is limited and poorly effective, making current research on terrain rendering focus on other techniques. Another similar feature, directly applicable to terrain rendering, is hardware tessellation [11]. This feature, that appeared in the shader model version 5, adds new stages in the graphic pipeline that allow to create new geometry "on the fly" more efficiently than the earlier shader stage, achieving high fidelity interactive terrain rendering [12].Hardware tessellation has become a de-facto standard nowadays.

On other hand, some very popular GPU post-processing techniques have been developed. These techniques enhance the appearance of the rendered image by applying changes at the pixel level using the fragment shader, without increasing the number of polygons. Two examples of these techniques are bump mapping [3] and parallax mapping [15]. Parallax mapping is actually an advance over bump mapping. It uses information included in a texture with height small variations of the reference surface (usually a flat surface) to modify the texture coordinates of the color texture applied to the model. In an intuitive way, parallax mapping corrects the coordinates of the color texture mapped on the reference surface, approximating them to its coordinates as if the full resolution surface had been ren-

dered as a polygonal surface instead. However, some features of parallax mapping prevent it from being used for terrain rendering. One of these features is that parallax mapping is designed for real surfaces with small variations on the reference surface (such as stone walls or floors). Nevertheless, a real terrain surface can show great height variability, since terrains typically have a fractal structure. Another feature is that the corrected coordinates become too large when the angle between the view ray and the reference surface is small, obtaining a wrong estimation. Since most of the applications that usually render a terrain surface (like computer games or driving simulators) place the virtual camera near the ground surface, parallax mapping cannot be used in the most extened kind of visualization applications.

In this paper, we present a new real-time terrain rendering approach which efficiently combines hardware tessellation and parallax mapping, making parallax mapping compatible with hardware tessellation and terrain rendering. The proposed approach uses a low resolution terrain model instead of a flat surface as the reference surface, and it determines "on the fly" the height variations on the terrain model, in order to reduce their size. The performance evaluation results show that the proposed scheme improves the performance of real-time terrain rendering applications in regard to the performance yielded when using hardware tessellation only.

The rest of the paper is organized as follows: section 2 describes the proposed approach for real-time terrain rendering. Next, section 3 shows the performance evaluation of the proposed model. Finally, section 4 shows some conclusion remarks.

2 A new real-time terrain rendering approach

We propose an approach that combines hardware tessellation and parallax mapping techniques. Both techniques are implemented in the graphics shaders. Following the OpenGL nomenclature, some code is placed in the vertex shader but the main code is implemented in the tessellation shaders (Tessellation Control Shader and Tessellation Evaluation Shader) and in the fragment shader. Figure 2 illustrates the location of these programmable shaders in the graphic pipeline. Hardware Tessellation appeared in the shader model version 5, with the introduction of two new programmable stages in the graphic pipeline: the Tessellation Control Shader (TCS) and the Tessellation Evaluation Shader (TES), and one fixed stage: the Tesselation Primitive Generator (TPG). In our approach, we have applied similar ideas to those in the NVIDIA whitepaper [12] to program these shaders.

2.1 Tessellation

Hardware tessellation requires that the base mesh of the tile is defined as a set of patch primitives. These primitives are general-purpose primitives defined from a set of vertices.

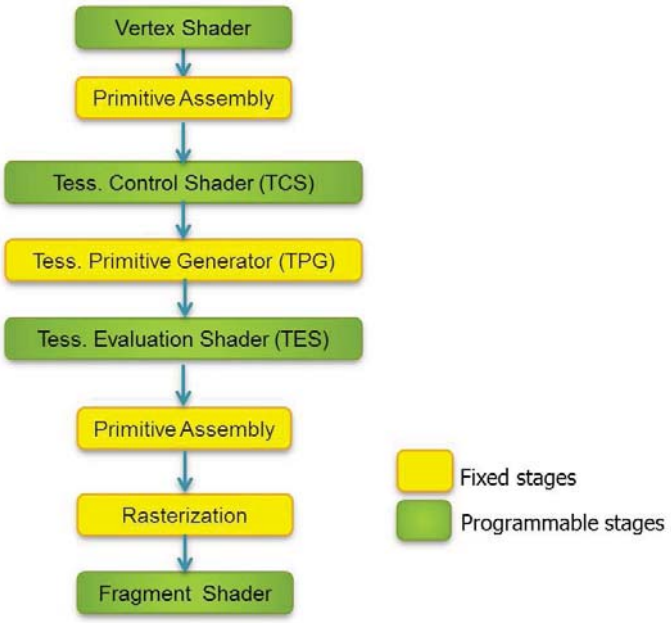


Figure 2: Graphic pipeline with the programmable stages

The maximum number of patch vertices is implementation-dependent, but the minimum number can be a single patch vertex.

On other hand, the tessellation technique uses heightmaps (textures with elevation information), in addition to color texture, in order to add geometry from a coarse mesh according to the distance to the observer, performing a smooth view-dependent level-of-detail representation. However, heightmaps typically do not fit into the GPU memory. In order to overcome these kind of hardware-imposed tessellation density limitations, the base geometry as well as the textures mapped onto it (color texture, heightmaps, normalmaps, etc.), are divided into rectangular regions (each one corresponding to one patch). At run-time, the textures of every new patch inside the camera view-frustum are sent to the GPU, where they are directly refined and rendered. When every patch is rendered with a view-dependent level-of-detail, some cracks (holes in the terrain surface) can appear along the edges. The solving of this problem requires the handling of patch boundaries resolution in an independent way of the inner mesh resolution, in addition to seams between heightmaps.

The Tessellation Control Shader (TCS) should determine how many times the patch must be subdivided, on the basis of the camera distance. It is also responsible for ensuring continuity across boundaries patches, forcing the shared edges between the patches to use the same level of tessellation. Figure 3 shows part of the code (using GLSL language) for

this shader. The control points and texture coordinates are passed unaltered to the next stage.

```
if (frustrumTest == true)
{
  gl_TessLevelOuter[0] = getTessLevel(distEdge0toCamera);
  gl_TessLevelOuter[1] = getTessLevel(distEdge1toCamera);
  gl_TessLevelOuter[2] = getTessLevel(distEdge2toCamera);
  gl_TessLevelOuter[3] = getTessLevel(distEdge3toCamera);

  gl_TessLevelInner[0] = getTessLevel(distCentertoCamera);
  gl_TessLevelInner[1] = getTessLevel(distCentertoCamera);
}
tcTexCoords[gl_InvocationID] = vTexCoords[gl_InvocationID];
gl_out[gl_InvocationID].gl_Position =
   gl_in[gl_InvocationID].gl_Position;
```

Figure 3: Tessellation Control Shader code

The TPG subdivides the patch based on the tessellation level values computed by the TCS, and the Tessellation Evaluation Shader (TES) computes the vertex values for each generated vertex. Figure 4 shows part of the code for this shader. It must be noted that we are using mipmapping with the heightmap texture to avoid aliasing problems.

```
vertexPos = computeVertexPosition(gl_TessCoord);
vertexPos.y = uHeightScale * texture(uHeightmap, tcTexCoords).r;

gl_Position = uModelViewProjectionMatrix * vertexPos;
```

Figure 4: Tessellation Evaluation Shader code

2.2 Parallax mapping

Parallax mapping uses information included in the heightmap texture to modify the texture coordinates of the color texture that is mapped on the reference surface (usually a flat mesh), approximating them to the coordinates of the full resolution surface. The process is similar to trace rays into the height field to obtain the texture coordinates of the visible point.

Simple parallax mapping [16] obtains the target texture coordinates (u', v') by the projection of the intersection point on the real surface (high resolution surface) assuming

that the heigh field $h(u, v)$ is near constant everywhere in the neighborhood of (u, v). This technique usually uses a flat mesh as reference surface. Figure 5 shows the process, and figure 6 shows the implementation of the Welsh method [17] for this type of reference surface.

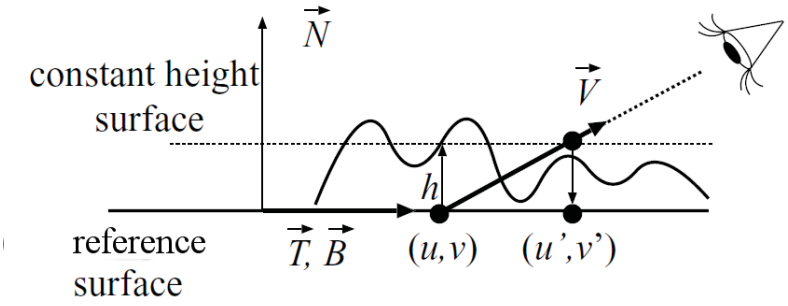


Figure 5: Finding the intersection point between the eye and the reference surface (flat mesh)

```
float hsb = texture(uHeightmap, teTexCoords.st).r;
newTexCoords = teTexCoords + hsb * eye.xz;
```

Figure 6: Implementation of the simple parallax mapping method for a flat reference surface

However, parallax mapping is not well-suited for terrain rendering when a flat reference surface is used, due to the great variability of terrain data that results in a wrong estimation of the corrected texture coordinates, specially when the angle between the view vector and the reference surface is small.

Nevertheless, we can avoid these problems using a reference surface closer to the real surface. Therefore, we use the mesh created in the tessellation stages of the graphic pipeline, which is a low resolution model of the terrain model, as reference surface in the parallax mapping algorithm. This process does not change the geometry of the surface rendered (which still remains the mesh created by the tessellation shaders), but the results are similar as if the full resolution surface had been rendered as a polygonal surface instead.

The heightmap (real surface) is defined in a planar domain, but we now need to compute height variations of the real surface from every planar triangle of the new reference surface in its normal vector direction. Therefore, we have modified the tessellation evaluation shader code to output the height hr and the normal vector direction nr at every point of the reference surface. Also, we have modified the parallax mapping algorithm to compute these variations in order to determine the new texture coordinates for the color texture, as shown in figure 7.

```
float hsb = (texture(uHeightmap, teTexCoords.st).r - hr) * nr.y;
newTexCoords = teTexCoords + hsb * eye.xz;
```

Figure 7: Implementation of the simple parallax mapping method for the reference surface created in the tessellation shaders

3 Performance evaluation

In this section, we analyze the performance of the proposed approach. Concretely, we compare the results obtained when exclusively using hardware tessellation with the results obtained when using hardware tessellation combined with parallax mapping. We have implemented a generic terrain visualization application to compare both strategies. This application performs a real-time fly over the terrain model using different visual quality parameters, and it allows taking screen captures of the rendered terrain from the user point-of-view and storing them for analysis.

The terrain database used in the tests is a subset of the Puget Sound database [18]. This database is usually used to test terrain visualization applications, due to its varied geography. The subset has been split in 4x8 regular tiles. Each one of them is a heightmap image of 1024x1024 pixels with a sample spacing resolution of 20 centimeters. This involves a total surface of 819 meters wide by 1638.5 meters long. The color textures have a resolution of 2048x2048 pixels. The total size of the test data is about 671 MB in its decompressed form. Tests were run on a PC-platform based on an Intel Core i7-4790 CPU processor, 12 GB of RAM and a NVIDIA GeForce GTX 650 GPU graphic card.

Both strategies have been tested using different levels of detail (tessellation factors), in order to study how they scale with the number of triangles. The terrain data set has 8x4 tiles and the underlying geometry has been set to 8 subdivisions, which translates into a 162 triangle mesh with the minimum tessellation factor, 1. The number of triangles scales linearly as the tessellation factor increases, until it reaches a total of 331776 triangles at the maximum tessellation factor, 64.

We have obtained the Mean Square Error (MSE) produced by the two considered techniques (Tessellation and Tessellation + Parallax) as a function of the number of triangles drawn in the scene, which in turn depends on the tessellation level, that is, the level of detail used by the tessellation algorithm for generating the mesh. Figure 8 shows these results, proving that the proposed technique yields a higher accuracy for all the levels of detail.

Figure 9 shows a quantitative measurement of the improvements achieved by the proposed technique, in regard to the performance obtained by tesselation. In this figure, each point represents a fixed number of triangles. For that number of triangles, the X-axis value is the MSE yielded by the tessellation technique, and the value in the Y-axis is the MSE yielded by the propoped technique (tessellation+Parallax). Since the slope of the plot is

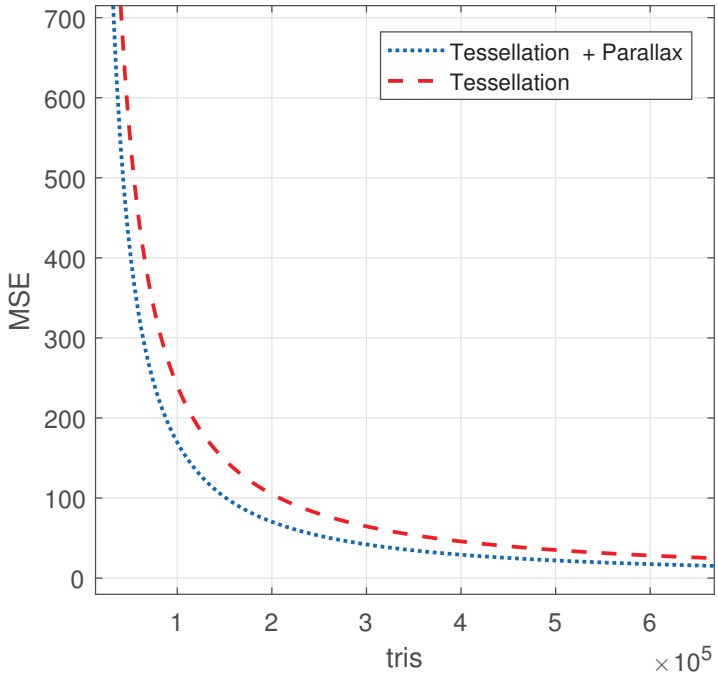


Figure 8: MSE as a function of the number of triangles in the mesh

around 75%, it means that the proposed technique yields and improvement of around 25%.

4 Conclusions

In this paper, we have proposed a new real-time terrain rendering approach which efficiently combines hardware tessellation and parallax mapping, making parallax mapping compatible with hardware tessellation and terrain rendering. The performance evaluation results show that the proposed scheme reduces the MSE of real-time terrain rendering applications in regard to the one yielded when exclusively using hardware tessellation. Therefore, we can conclude that the proposed technique not only makes possible the use of the de-facto standard of hardware tessellation in these applications, but it also improves the rendering performance

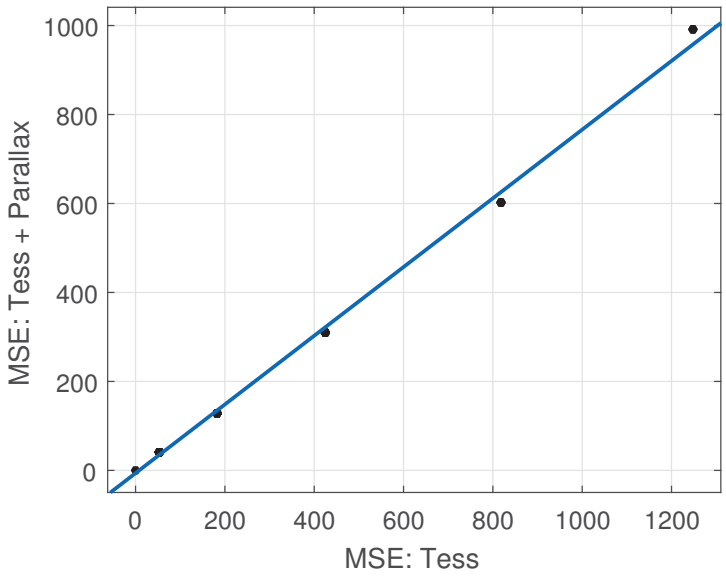


Figure 9: Proportion between the MSE yielded by each technique.

Acknowledgements

This work has been supported by Spanish MINECO and EU FEDER funds under grant TIN2015-66972-C5-5-R.

References

[1] *Microsoft, 2011. Flight simulator home page*, Retrieved April 27, 2016 from http://www.microsoft.com/games/fsinsider

[2] *Square Enix Co., 2015. Final Fantasy XIV home page*, Retrieved May 1, 2015 from http://www.finalfantasyxiv.com/

[3] James F. Blinn, *Simulation of wrinkled surfaces*, Proc. of 5th conference on Computer Graphics (SIGGRAPH '78). ACM, New York, NY, USA, (1978) 286-292.

[4] R. Pajarola, E. Gobbetti, *Survey of semi-regular multiresolution models for interactive terrain rendering*, The Visual Computer **23**.8 (2007) 583–605.

[5] F. Losasso, H. Hoppe, *Geometry clipmaps: terrain rendering using nested regular grids*, ACM Transactions on Graphics (TOG) **23**.3 (2004) 769–776

[6] J. Schneider, R. Westermann, *GPU-friendly high-quality terrain rendering*, Journal of WSCG **14**.1–3 (2006) 49–56

[7] Y. Livny, N. Sokolovsky, T. Grinshpoun, J. El-Sana, *A GPU persistent grid mapping for terrain rendering*, The Visual Computer **24**.2 (2008) 139–153.

[8] Y. Livny, Z. Kogan, J. El-Sana, *Seamless patches for GPU-based terrain rendering*, The Visual Computer **25**.3 (2009) 197–208.

[9] C. Dick, J. Krger, R. Westermann, *GPU ray-casting for scalable terrain rendering*, Proceedings of EUROGRAPHICS - Areas Papers **50** (2009) 43–50.

[10] D. Blythe, *The Direct3D 10 system*, ACM Transactions on Graphics (TOG) **25**.3 (2006) 724–734.

[11] J. Zink, M. Pettineo, J. Hoxley, *Practical rendering and computation with Direct3D 11*, CRC Press, 2011.

[12] I. Cantlay, *Directx 11 terrain tessellation*, Nvidia whitepaper **8** (2011) 11.

[13] Olanda, R., Pérez, M., Orduña, J. M. (2013), *Hybrid P2P schemes for remote terrain interactive visualization systems*, Future Gener. Comput. Syst. 29(2013), 1522–1532.

[14] Olanda, R., Pérez, M., Orduña, J. M., Rueda, S. (2014)., *Terrain data compression using wavelet-tiled pyramids for online 3D terrain Visualization.*, Int. Journal of Geographical Information Science 28(2014), No. 2, 407–425.

[15] L. SzirmayKalos, T. Umenhoffer, *Displacement Mapping on the GPUState of the Art*, Computer Graphics Forum **27**.6 (2008) 1567–1592.

[16] T. Kaneko, T. Takahei, M. Inami, N. Kawakami, Y. Yanagida, T. Maeda, S. Tachi, *Detailed shape representation with parallax mapping*, Proceedings of ICAT (2001) 205–208.

[17] Welsh, Terry., *Parallax mapping with offset limiting: A perpixel approximation of uneven surfaces*, Infiscape Corporation (2004) 1–9.

[18] USGS and The University of Washington, *Puget Sound Terrain*, Retrieved April 15, 2016 from `http://www.cc.gatech.edu/projects/large_models/ps.html`

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Controlling Oscillations of a Nonlinear Hanging String with a Tip Mass

G. González-Santos[1] and C. Vargas-Jarillo[2]

[1] *Departamento de Matemáticas, ESFM-IPN,*
Unidad Profesional Zacatenco, México D. F. 07738.

[2] *Departamento de Control Automático, CINVESTAV-IPN*
A.P. 14-740 México D.F. 07000.

emails: gsantos@esfm.ipn.mx, cvargas@ctrl.cinvestav.mx

Abstract

To reduce the amplitude of the oscillations of a nonlinear vertical heavy string with a tip mass at the end free is an important mechanical problem. We analyze two ways to attain this goal. One of them is to change systematically the length of the string. The second one is to replace some part of the elastic string by viscoelastic components. The string is modeled by an array of N masses connected by N massless linear elastic springs. At the bottom end the string has a tip mass. The interacting force between the particles is the classical Hooke's force and the nonlinearity is due to the geometry the problem.

Key words: Nonlinear vibrations, Elastic string, Hanging string, Variable length pendulum, Viscoelastic.
MSC 2000: 65L05, 70F10

1 Introduction

There are several mechanical problems where to reduce the amplitude and the velocity of the free end of a vertical string with a tip mass is important. For instance, the amplitude of a simple pendulum can be reduced by systematically varying its length, [10]-[12]. Other example, is the control of the vibrations of the hanging string with a tip mass by replacing one o several springs by a Kelvin's unit. The study of the hanging chain with a tip mass has been considered by Sujith and Hodges [13] who derived an exact frequency relation.

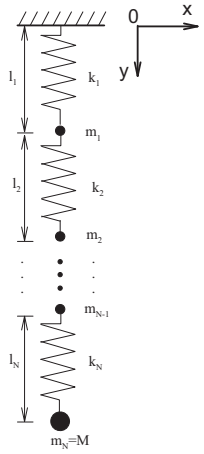


Figure 1: Discrete model of a hanging string with a tip mass.

This study has been extended and tested experimentally by Deschaine and Suits [2]. Some numerical studies has been done by C. Y. Wang, [14].

In this paper we use a discrete model, [3]-[7], [9], to study the vertical heavy string with a tip mass. The description of the discrete model is contained in the second section. The third section is dedicated to the variable length pendulum and in the last one we analyze how the amplitude oscillations of a heavy vertical string can be reduced by replacing springs by Kelvin elements.

2 Model formulation

The discrete model of a hanging string with a tip mass consists of N particles, $P_1, P_2, ..., P_N$ with masses $m_1, m_2, ..., m_N$ respectively, joined by N Hookean massless springs. The ith particle is connected to the $(i+1)th$ particle by a ith spring with stiff constant k_i. The initial lengths of the springs are $l_1, l_2, ..., l_N$ ($L = l_1 + l_2 + ... + l_N$). The upper end of the first spring is fix and the lower end of the last spring is connected to the tip mass ($M = m_N$). The discrete model is shown in Fig. 1. The variables at time t are:

$\mathbf{r}_i(t) = (x_i(t), y_i(t))$	Coordinate of the particle P_i,
$\mathbf{r}_{i,j}(t)$	Vector $\mathbf{r}_i - \mathbf{r}_j$,
$r_{i,j}(t)$	Norm of the vector $\mathbf{r}_{i,j}$,
$\mathbf{v}_i(t)$	Velocity of the particle P_i,
$\mathbf{a}_i(t)$	Aceleration of the particle P_i,
$\mathbf{F}_i^*(t)$	Force acting on P_i due to nearest neighbor particles,
$\mathbf{f}_i(t)$	Long range force acting on particle P_i (gravity), and
$\mathbf{F}_i(t)$	Total force acting on P_i
	for $i = 1, 2, ..., N$.

The force $\mathbf{F}_i^*$ exerted on the particle P_i by the springs i and $i+1$ is given by:

$$\mathbf{F}_i^* = -\left[k_i(r_{i,i-1} - l_i)\right] \frac{\mathbf{r}_{i,i-1}}{r_{i,i-1}} + \left[k_{i+1}(r_{i,i+1} - l_{i+1})\right] \frac{\mathbf{r}_{i,i+1}}{r_{i,i+1}}, \tag{1}$$

where $r_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$ is the Euclidian distance between the particles P_i and P_j. This expression introduces the nonlinearity in our problem since we are allowing vibrations in two dimensions. In contrast to the usual model where the particles are constrained to move only along the horizontal axes.

The total force acting upon the particle P_i is

$$\mathbf{F}_i = \mathbf{F}_i^* + \mathbf{f}_i.$$

The acceleration of P_i at time t is related to the force by a discrete Newton's Law:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i(t), i = 1, 2, ..., N \tag{2}$$

Thus (2) is a system of 2N second order differential equations. In general this system can not be solved analytically from initial positions and velocities, therefore it must be solved numerically.

Since the velocity, at time t, of the particle P_i is $\mathbf{v}_i$ we can determine the kinetic energy of the string by

$$T(t) = \frac{1}{2} \sum_{i=1}^{N} m_i \parallel \mathbf{v_i} \parallel^2 .$$

The potential energy $V(t)$ is found by considering the gravity effect and the increment or decrement of length of each spring. The ith spring increases or decreases its length from l_i to $r_{i,i-1}$. Therefore, we have done an amount of work $\frac{1}{2}k_i(r_{i,i-1} - l_i)^2$. Summing up for all the springs, we obtain the potential energy of the string at time t:

$$V(t) = \sum_{i=1}^{N} \frac{1}{2} k_i (r_{i,i-1} - l_i)^2 + m_i g h_i.$$

 ISBN: 978-84-608-6082-2

where h_i is the height of the ith particle relative to the equilibrium position of the $N-th$ particle (y_N^{eq}) which is obtained from (3).

2.1 Equilibrium position

The equilibrium position of the particles of the string, $\mathbf{r}_i^{eq} = (0, y_i^{eq})$, $i = 1, 2, ..., N$, is obtained by taking the x component equal to zero and the right hand side of (2) equal to zero. The y-component of (2) becomes a linear system, (3), which is solved for each initial condition.

$$\mathbf{K}\mathbf{y}^{eq} = -(\mathbf{T}\mathbf{I} + g\mathbf{m}), \tag{3}$$

where

$$\mathbf{K} = \begin{pmatrix} -(k_1+k_2) & k_2 & & \\ k_2 & -(k_2+k_3) & k_3 & \\ & \ddots & \ddots & k_N \\ & & k_N & -k_N \end{pmatrix},$$

$$\mathbf{T} = \begin{pmatrix} k_1 & -k_2 & & \\ & k_2 & -k_3 & \\ & & \ddots & -k_N \\ & & & k_N \end{pmatrix},$$

and $\mathbf{m} = [m_1, m_2, ..., m_N]^t, \mathbf{l} = [l_1, l_2, ..., l_N]^t$

3 Variable length pendulum

Controlling the angular oscillations of a simple pendulum is related with a child's swing. It is well known that to swing a swing one must be crouch near to equilibrium position and straighten up at the extreme position to increment the period of oscillation. One way to control the pendulum oscillations is by adjusting the length of the cable $r(t)$, [10, 11, 12]. The motion equation of the simple pendulum with variable length is given by

$$mr^2\ddot{\theta} + 2mr\dot{r}\dot{\theta} = -mrgsin(\theta),$$

where the term $F_c = 2mr\dot{r}\dot{\theta}$ is known as the inertia force associated with the the sliding motion of the mass. This term can be use to increase o decrease the amplitude of the angular oscillations. As a simple mass-spring system with damping, when $\dot{r} > 0$ the F_c

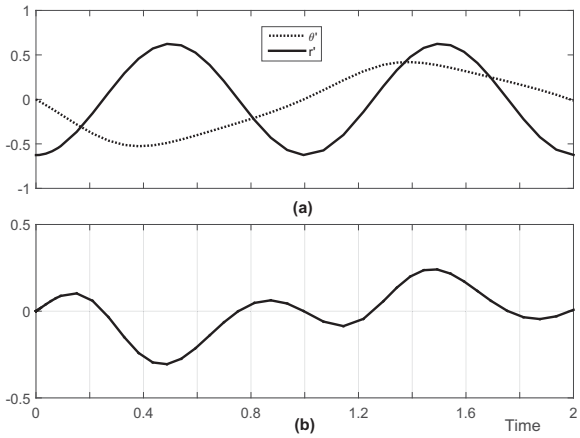


Figure 2: a) Angular velocity of the simple pendulum along a period (dotted line) and the temporal derivative of pendulum length that increases the F_c force effect (continuum line). b) The evolution of the product $\dot{r}\dot{\theta}$.

force is opposite to the rotation of the pendulum decreasing the amplitude of the oscillations. On other hand when $\dot{r} < 0$ the F_c force impels the pendulum rotation. One way to reduce the amplitude of the oscillations is by increasing and decreasing the pendulum length appropriately.

Fig. 2(a) shows the angular velocity ($\dot{\theta}$) of the pendulum during a oscillation period. At $t = 0$ the mass of the pendulum is θ radians far from the vertical and its velocity is zero. The maximum angular velocity, in magnitude, is attained when the mass move toward o away from the equilibrium position. If the pendulum length is increased when the magnitude of the angular velocity is hight then the F_c force is maximum and this maximize the damping effect. The length of the pendulum should be reduced when the angular velocity is small. One way to attain this goal is to vary the length of the pendulum periodically with a frequency twice the pendulum natural frequency (w). In Fig. 2(b) we show $\dot{r}\dot{\theta}$ of a simple periodic function, $r(t) = L \quad \Delta L sin(2wt)$, which reduces substantially the amplitude of the pendulum oscillations.

After the above strategy has been applied the period of the pendulum changes and it requires a period estimation before to applying the strategy to the next period. An example of the pendulum evolution over time is shown in Fig. 3 and Fig. 4. The angle between the vertical and the pendulum decreases to zero as the potential and kinetic energy also does .

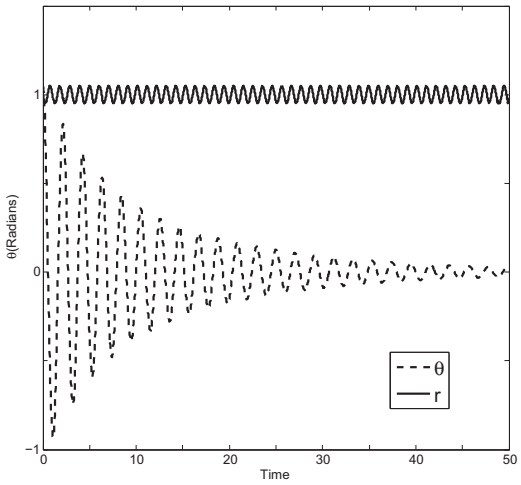


Figure 3: Angular displacement evolution when the length of the pendulum is equal to $r(t) = L - \Delta L sin(2wt)$, $\Delta L = 0.1$.

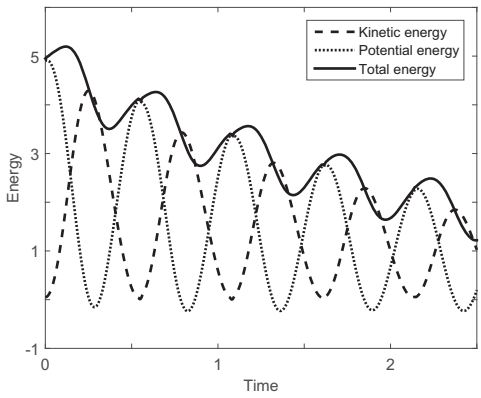


Figure 4: Kinetic, Potential and total energy of the pendulum evolution when the length of the pendulum is equal to $r(t) = L - \Delta L sin(2wt)$, $\Delta L = 0.1$

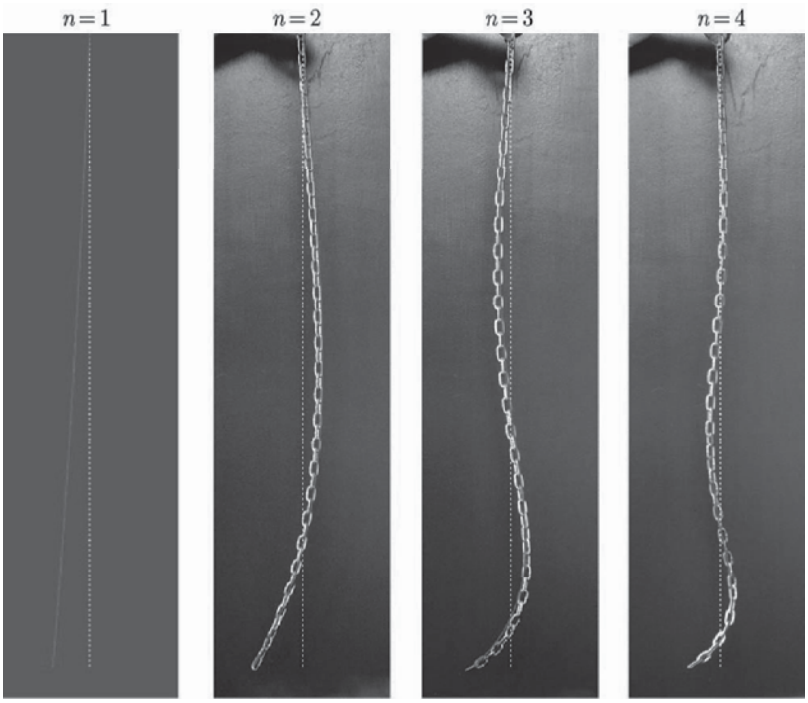


Figure 5: D. Young [15]. First four vibrational modes of the hanging chain. The analytical solution is shown in red.

4 Hanging string with a tip mass

We use the discrete model of the section 2 to analyze the hanging string with a tip mass behavior. The mass of the string is $m = m_1 + ... + m_{N-1} = 1$ with $m_1 = ... = m_{N-1}$. The mass of the tip is $m_N = M$. First we analyze the hanging string behavior when a initial velocity is applied at the free end, the string density is constant, $m_1 = ... = m_N$, the length of the string is $L = 1$, and its stiffness is also constant, $k_1 = k_2 = ... = k_N = 1000$. The gravity is $g = 0.1$ and $N = 32$ the number of particles. The initial conditions for the discrete model are given by:

$$\begin{aligned} \mathbf{r}_i(0) &= (0, y_i^{eq}), i = 1, 2, ..., N, \\ \mathbf{v}_i(0) &= (0, 0), i = 1, 2, ..., N-1, and \\ \mathbf{v}_N(0) &= (\beta, 0), \beta = 2.5E-1. \end{aligned}$$

The initial impulse produces a sudden displacement to the right of the free end and

 ISBN: 978-84-608-6082-2

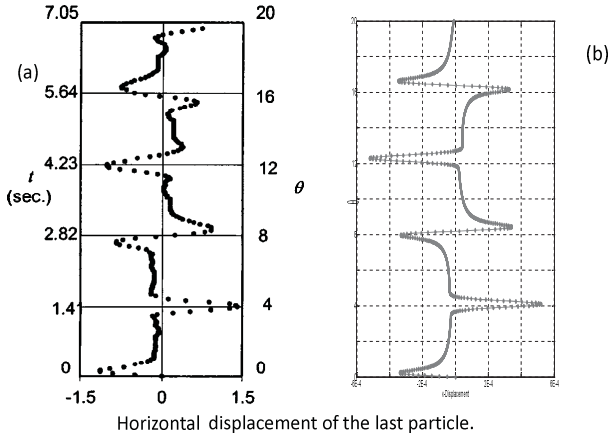


Figure 6: (a) H. Bailey, [1] . Experimental data from a movie showing the position of the last particle. (b) Discrete model.

then it returns quickly near to the equilibrium position, without swinging to the left and remains there for a while. Next a sudden jump to the left of the free end occurs and returns again near to the equilibrium position. Immediately after a whip of the free end it oscillates quickly from left to the right and finish again near to the equilibrium position. After reaching this position the free end suddenly jumps to the right and returns again near the equilibrium position. Finally the free end arrives near to the initial condition (zero displacement and velocity β).

Some experiments, with this kind of strings, had been carried out for D. Young [15], he shows a perfect matching between the first four vibrational modes of a hanging string and its analytical approximation and it is shown in Fig. 5. The motion of the free end of the string was obtained experimentally by H. Bailey in [1] and a comparison with the numerical results obtained from own discrete model are shown in Fig.6. Both results match satisfactorily.

In general the string behavior depends mainly on the quotient $q = M/m$, when the gravity and its stiffness are constants. When this quotient is bigger than the previous case the displacement of the tip mass is described for a Lissajous curve. The tip mass behavior for $M/m = 1, 10, 1000$ and 1000 is shown in Figs. 7(a)-7(d). As the quotient q increases the length of the string at equilibrium position increases and the frequency of the tip mass decreases. While the x-displacement is symmetric around the equilibrium position the y component is not. The oscillations of the hight frequency are reduced as the quotient increases. For instance, the Fig. 7 shows the tip evolution when $M/m = 1$, it has a wide range of oscillations that are reduced when q is bigger, as shown in Figs. 7(b)-7(d).

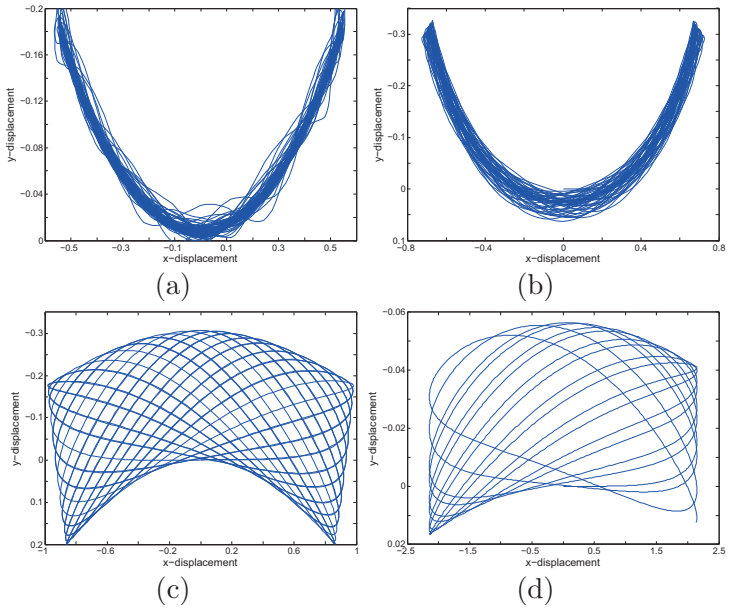


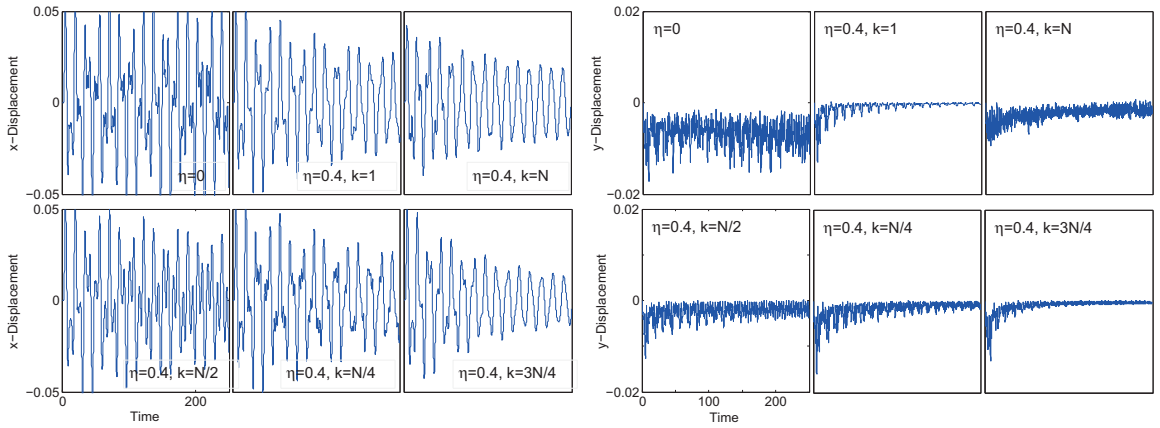
Figure 7: Evolution of the tip mass for (a) $M/m = 1$, (b) $M/m = 10$, (c) $M/m = 100$, (d) $M/m = 1000$.

5 Control of the tip mass amplitude

In some applications it is important to reduce the amplitude of the tip mass as soon as possible. To reduce the amplitude and the velocity of the free end we replace one o several springs by Kelvin units. A Kelvin unit is a mechanical device formed by a spring and a damper in parallel. When the $i-th$ spring is replaced by a Kelvin unit the equation (2) of the model is replaced by

$$\mathbf{F}_i^* = -[k_i(r_{i-1,i} - l_i) + \eta_i \dot{r}_{i-1,i}]\frac{\mathbf{r}_{i-1,i}}{r_{i-1,i}} + [k_{i+1}(r_{i,i+1} - l_i) + \eta_{i+1}\dot{r}_{i,i+1}]\frac{\mathbf{r}_{i,i+1}}{r_{i,i+1}}, \qquad (4)$$

where $\dot{r}(t)$ denotes the temporal derivative of $r(t)$ and η_i is the viscosity coefficient of the Kelvin unit. The system (2) with $\mathbf{F}_i^*$ as in (4) is solved numerically by using the fortran subroutines DDRIV2 [8]. The relative accuracy in the all solution components was taken equal to $1E-6$.



(a) (b)

Figure 8: Evolution of the (a) x-component and (b)y-component of the displacement of the free end when a spring is replaced for a Kelvin unit at different positions of the string. $k = 1$ (fix end), $k = N$ (free end).

5.1 One replacement

The evolution of the x and y components of the tip displacement relative to the equilibrium position are shown at the top-left in Fig.8(a-b) respectively. The amplitude of x-displacement is at least three times bigger than y-displacement. Both displacements decrease when one spring is replaced for one Kelvin unit. From the same figure it can be observed that when one Kelvin unit is put at position $k = 3N/4$, is more effective in slowing down the oscillations. The same figures show that the y-component tend to zero faster than the x-component. How fast tends to zero depends on the unit Kelvin position and its parameter η. Besides the position of the Kelvin unit has a different damping effect over each component of the displacement. When the Kelvin unit is located near the fix end the y-displacement tends to the equilibrium position in the fastest way. However the x-component is more affected by the Kelvin unit near the free end ($k = 3N/4$). In all cases η was 0.4.

5.2 Two replacements

Now two springs are replaced by two kelvin units, we keep the position $3N/4$ which gave the best results when one spring is replaced. According with the Fig. 9(a-b) when the second Kelvin unit is located at the position $k = N$ the x-oscillations of the free end decreases and are minimum among the other positions after a fix time. The decrement of y-component of the displacement is almost independent of the second Kelvin unit position. Although the high frequency oscillations are reduced when the second Kelvin unit is at $k = N/2$.

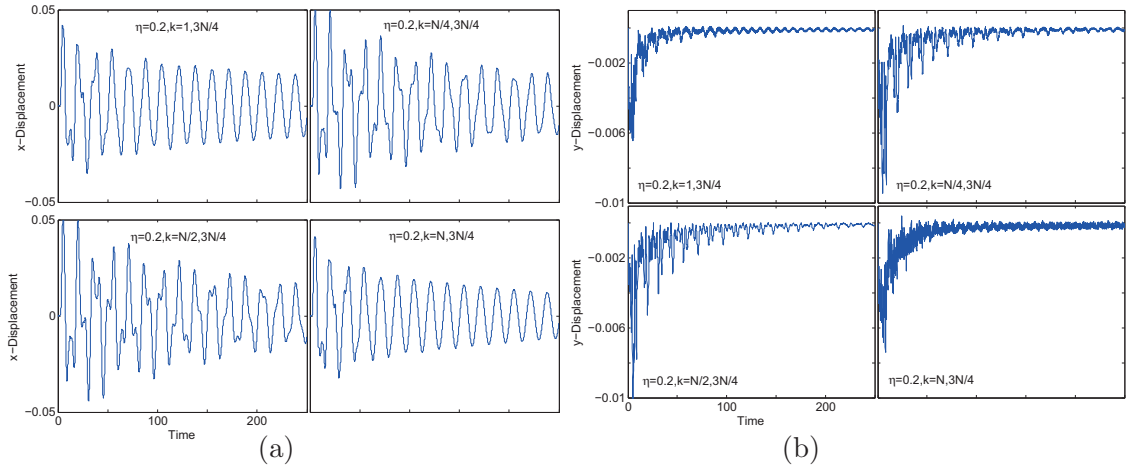


Figure 9: Evolution of the (a) x-component and (b) y-component of the displacement of the free end when two springs are replaced for Kelvin units at different positions of the string. $k = 1$ (fix end), $k = N$ (free end).

6 Conclusions

According with our numerical experiments, to replace one o several spring by Kelvin units is an effective way to slow down the amplitude oscillations of the tip mass of a hanging string. The replacement procedure is more efficient for the y than the x-component. On other hand, variable length strategy can reduce the angular displacement and velocity of a simple pendulum. The frequency of the length cable variation must be the double of the natural frequency of the pendulum.

References

[1] H. Bailey, *Motion of a hanging chain after the free end is given an initial velocity.* Am. J. Phys. 68(8), 2000.

[2] J. S. Deschine and B. H. Suits, *The hanging cord with a real tip mass.* Eur. J. Phy. 29, Pag. 1211-1222, 2008.

[3] G. González-Santos and C. Vargas-Jarillo *A numerical study of the elastic hanging string with a tip mass* Proceeding othe 12th Conference of Computational and Mathematical Methods in Science and Engineering, CMMSE 2012.

[4] G. González-Santos and C. Vargas-Jarillo *A Three dimensional model of a nonlinear hanging string with a tip mass* Proceeding othe 14th Conference of Computational and Mathematical Methods in Science and Engineering, CMMSE 2014.

[5] D. Greenspan, *Computer simulation of transverse string vibrations*. BIT 11, Pag. 339-408, 1971.

[6] D. Greenspan, *Discrete models*. Addison-Wesley, 1973.

[7] D. Greenspan, *N-Body problems and models*.World Scientific, 2004.

[8] D. Kahaner, C. Moler and S. Nash, *Numerical methods and software*. Prentice Hall, 1989.

[9] J. P. McCreesh, T. L. Goodfellow, and A. H. Seville *Vibrations of a hanging chain of discrete links*. Am. J. Phys. Vol. 43, No. 7, 1975.

[10] A. P. Seyranian, *The swing: parametric resonance*. J. Applied Math. and Mech. Vol. 68, pp. 757-764. 2004.

[11] A. P. Seyranian, A. O. Belyakov, *Swing dynamics*. Doklady Physics, Vol. 53, No. 7, pp. 338-394. 2008.

[12] D. S. D. Stilling, W. Szyszkowski, *Controlling angular oscilations through mass reconfiguration: a variable length pendulum case*. Int. J. of Non-Linear Mechanics 37, pp. 89-99. 2002.

[13] R. I. Sujith D. H. Hodges, *Exact solution of the free vibration of a hanging cord with tip mass*. J. Sound Vibr. No. 179, pp. 359-361. 1995.

[14] C. Y. Wang and C. M. Wang, *Exact solution for vibration of a vertical heavy string with a tip mass*. The IES Journal Part A: Civil and Structural Engineering, Vol. 3. No. 4, Pag. 278-281, 2010

[15] D. Yong, *Strings, chains and ropes*. SIAM REVIEW. Vol. 48, No. 4, pp. 771-781, 2006.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Thwarting randomness reveals in group key agreement

María Isabel González Vasco[1], Ángel L. Pérez del Pozo[1] and Adriana Suárez Corona[2]

[1] *Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, Spain*

[2] *Research Institute of Applied Sciences in Cybersecurity, RIASC, Departamento de Matemáticas, Universidad de León, Spain*

emails: mariaisabel.vasco@urjc.es, angel.perez@urjc.es, asuac@unileon.es

Abstract

When a group key exchange protocol is executed, the session key is typically extracted from two types of secrets; long-term keys (for authentication) and freshly generated (often random) values. The leakage of this latter so-called *ephemeral keys* has been extensively analyzed in the 2-party case, yet very few works are concerned with it in the group setting. We provide a generic construction that is strongly secure, meaning that the attacker is allowed to learn both long-term and ephemeral keys (but not both from the same participant because this would trivially disclose the session key). Our construction can be seen as a compiler, in the sense that it builds on a 2-party key exchange protocol which is strongly secure and transforms it into a strongly secure group key exchange protocol by adding only one extra round of communication. When applied to an existing 2-party protocol from Bergsma et al., the result is a 2-round group key exchange protocol which is strongly secure in the standard model, thus yielding the first construction with this property.

Key words: group key exchange, compiler, strong security, ephemeral keys

1 Introduction

Group key establishment (GKE) protocols are a fundamental cryptographic building block allowing $n \geq 2$ participants to agree upon a common secret key. It is usually assumed that these participants hold both *long-term* secrets, which are typically used for authentication and *ephemeral* secrets,

which are session-specific randomly generated values that provide enough entropy for the key to be indistinguishable from random in some sense.

The way to define and handle key privacy is highly dependent on the amount of information the adversary is supposed to be able to derive of the two types of "secrets" described above. In the literature, leakage of ephemeral secrets is often modeled through a RevealState oracle, which when invoked by the adversary outputs either *ephemeral keys* as described above or a larger set containing them, typically referred to as the *full state* of the attacked user. Unfortunately, as first pointed out by Cremers in [8], the meaning of *full state* is scarcely defined within the security model and often the output of the corresponding oracle calls is only made explicit when proving particular protocols secure. Generally speaking, ephemeral key leakage refers to the exposure of user-generated fresh randomness, while full state compromise involves in addition other values computed/stored by the user —yet never any long-term keys.

Previous work. *Strong security* for GKE protocols was first considered in [6]. This setting brings into play a new type of adversary who may gain access to ephemeral keys in addition to long-term secrets. However, many proposals dealing with this type of leakage [6, 5, 10, 11] actually assume that the adversary cannot access any ephemeral secret *of the attacked session*.

In order to subsume a wider class of attacks, other works have removed this restriction excluding only reveals of *both* the ephemeral and long-term secrets of the same user (as, in this case, the session key would be trivially disclosed). Some examples of secure proposals in this type of models are the NAXOS protocol [12] in the 2-party setting and [13, 9] for the case of 3 users. In the general multi-user group setting Zhao et al. [15] modify a protocol by Bohli et al. [4] to obtain strong security; nevertheless this proposal was found flawed in [7] where an improvement was proposed.

Many of these previous works have in common that the access to the ephemeral secrets is modelled granting the adversary a RevealState oracle, which, when queried, outputs the contents of a variable state linked to the execution. As pointed out by Cremers [8], the security in these models is highly dependant on how the state variable is defined for each concrete protocol; in addition Cremers shows that the NAXOS protocol, proven secure in the model with a different formalism (namely, defining a so-called RevealEphemeralKey oracle), is insecure when certain state reveals are allowed. Also in this line, in a recent work from PKC2015, Bergsma et al. [3] present a generic 1-round 2-party key exchange construction in the standard model. The authors also propose a strong security model which builds on previous ones and captures both perfect forward secrecy and ephemeral secrets leakage. The latter is modeled by a RevealRand oracle which outputs the local randomness selected by the user in a protocol execution.

Our contributions. We propose a security model for GKE capturing the leakage of ephemeral secrets even within the attacked session. To avoid any ambiguity we define, in the line of [3] in the 2-party setting, a variable rand that stores, for each instance of a participant, all the session values that cannot be computed from long-term secret keys or other values received/computed previously in the session. Typically these values are chosen uniformly at random from a prescribed set, therefore the name of the variable. The adversary is given access to an oracle RevealRand which outputs the

values stored in rand when queried. The strength of our security model is comparable to those in [15] or [7] (although they are not fully comparable as evidenced by Cremers [8]).

In addition, the main contribution of this work is a generic protocol or *compiler* that, building on any strongly secure 2-party authenticated key exchange (AKE) protocol, produces a group AKE which is strongly secure in our model, by adding only one round of communication. The construction is in the standard model. Further, we highlight that:

- Our constuction is the first to explicitly take into account the randomness used for authentication. We do so by expliciting that the nonces involved in any signature produced are part of the rand variable linked to the signing instance, and therefore allowing the adversary to obtain these values. We provide a security notion for signatures withstanding randomness reveal and discuss how to achieve it. This improves previous works, such us [7] where the random values for the signature are supposed to be protected or absent (by using a deterministic signature).
- When instantiated with a 1-round 2-party protocol strongly secure in the standard model, for instance [3], our compiler produces a 2-round GKE which is strongly secure in the standard model. As far as we know there is no other construction in the literature with this property.

2 Security Model

Our security model is a modification of the model of Bohli et al [4] and quite similar in strength and security requirements to those in [15] and [7]. It differs from previous ones in that it treats separately the session state (state) and the session ephemeral keys (rand) from a given session, following the rationale from [8].

Protocol instances. Users are modeled as probabilistic polynomial time (ppt) Turing machines. Each user from a set $\mathcal{U}$ of possible participants may execute a polynomial number of protocol *instances* in parallel. To refer to instance s_i of a user $U_i \in \mathcal{U}$ we use the notation $\Pi_i^{s_i}$ $(i, s_i \in \mathbb{N})$. To each instance we assign seven variables, informally described next:

- $\mathsf{used}_i^{s_i}$ indicates whether this instance is or has been used for a protocol run;
- $\mathsf{state}_i^{s_i}$ keeps the internal state of the Turing machine that executes the protocol;
- $\mathsf{rand}_i^{s_i}$ keeps the session-specific atomic secret values –typically values generated uniformly at random — which will be referred to as *ephemeral keys*. More precisely, these are any values that cannot be computed from long-term secret keys or other values received/computed previously in the session;
- $\mathsf{term}_i^{s_i}$ shows if the execution has terminated;
- $\mathsf{sid}_i^{s_i}$ denotes a session identifier;

 ISBN: 978-84-608-6082-2

- $\mathsf{pid}_i^{s_i}$ stores the set of identities of those users that $\Pi_i^{s_i}$ aims at establishing a key with—including U_i himself;
- $\mathsf{acc}_i^{s_i}$ indicates if the user accepted the session key;
- $\mathsf{sk}_i^{s_i}$ stores the session key once it is accepted by $\Pi_i^{s_i}$.

Communication network. Arbitrary point-to-point connections among the users are assumed to be available. The network is non-private, however, and fully asynchronous. More specifically, it is controlled by the adversary, who may delay, insert and delete messages at will.

Adversarial capabilities. We restrict to ppt adversaries. The capabilities of an adversary $\mathcal{A}$ are made explicit through a number of *oracles* allowing $\mathcal{A}$ to communicate with protocol instances run by the users:

- $\mathsf{Send}(U_i, s_i, M)$ This sends message M to the instance $\Pi_i^{s_i}$ and returns the reply generated by this instance. If $\mathcal{A}$ queries this oracle with an unused instance $\Pi_i^{s_i}$ and $M \subseteq \mathcal{P}$, a set of identities of principals, the $\mathsf{used}_i^{s_i}$-flag is set, $\mathsf{pid}_i^{s_i}$ initialized with $\mathsf{pid}_i^{s_i} := \{U_i\} \cup M$, and the initial protocol message of $\Pi_i^{s_i}$ is returned.
- $\mathsf{Execute}(\{\Pi_{u_1}^{s_{u_1}}, \ldots, \Pi_{u_\mu}^{s_{u_\mu}}\})$ This executes a complete protocol run among the specified unused instances of the respective users. The adversary obtains a transcript of all messages sent over the network. A query to the Execute oracle is supposed to reflect a passive eavesdropping.
- $\mathsf{Reveal}(U_i, s_i)$ Yields the session key $sk_i^{s_i}$ along with the session identifier $\mathsf{sid}_i^{s_i}$.
- $\mathsf{Test}(U_i, s_i)$ Provided that the session key is defined (i. e. $\mathsf{acc}_i^{s_i} = \mathtt{true}$ and $\mathsf{sk}_i^{s_i} \neq \perp$) and instance $\Pi_i^{s_i}$ is fresh (we define freshness later on), $\mathcal{A}$ can execute this oracle query at any time when being activated. Then, the session key $\mathsf{sk}_i^{s_i}$ is returned if $b = 0$ and a uniformly chosen random session key is returned if $b = 1$, where b is a hidden bit chosen at random prior to the first call. Namely, an arbitrary number of Test queries is allowed for the adversary $\mathcal{A}$, but once the Test oracle returned a value for an instance $\Pi_i^{s_i}$, it will return the same value for all instances partnered (also defined later) with $\Pi_i^{s_i}$. [1]
- $\mathsf{RevealRand}(U_i, s_i)$ This oracle returns the value stored in $\mathsf{rand}_i^{s_i}$.
- $\mathsf{Corrupt}(U_i)$ This oracle returns the long term key hold by U_i. We say user U_i is *honest* or *uncorrupted* if $\mathcal{A}$ has not made a call $\mathsf{Corrupt}(U_i)$ previously.

[1] This is the so-called Real or Random model, which can be proven equivalent to the usual model allowing for only one Test query with a loss of a factor m in the reduction, m being the number of involved protocol instances. See, for instance [1, 2].

 ISBN: 978-84-608-6082-2

We aim at two basic goals for our protocol: correctness and strong security. A protocol is correct if all users involved in an execution in the presence of a passive adversary compute the same session key. Our notion of strong security ensures key privacy in the presence of an active adversary which is given access to all the oracles we have described. Before formally defining correctness and strong security, we introduce *partnering* and *freshness*, to express which instances are associated in a common protocol session and limit when the adversary is allowed to call the Test oracle.

Partnering. We refer to instances $\Pi_i^{s_i}$, $\Pi_j^{s_j}$ as being *partnered* if $\mathsf{sid}_i^{s_i} = \mathsf{sid}_j^{s_j}$, $\mathsf{pid}_i^{s_i} = \mathsf{pid}_j^{s_j}$ and $\mathsf{acc}_i^{s_i} = \mathsf{acc}_j^{s_j} = \mathsf{true}$.

Freshness. A Test-query should only be allowed to those instances holding a key that is not for trivial reasons known to the adversary. To this aim, an instance $\Pi_i^{s_i}$ is called *fresh* if

- $\mathsf{acc}_i^{s_i} = \mathsf{true}$
- $\mathcal{A}$ never called $\mathsf{Reveal}(U_j, s_j)$ with $\Pi_i^{s_i}$ and $\Pi_j^{s_j}$ being partnered.
- If $\Pi_i^{s_i}$ and $\Pi_j^{s_j}$ are partnered and $\mathcal{A}$ called $\mathsf{Corrupt}(U_j)$, then any message sent to $\Pi_i^{s_i}$ on behalf of $\Pi_j^{s_j}$ must indeed come from $\Pi_j^{s_j}$ intended to $\Pi_i^{s_i}$.
- $\mathcal{A}$ never called both $\mathsf{Corrupt}(U_j)$ and $\mathsf{RevealRand}(U_j, s_j)$ with $\Pi_i^{s_i}$ and $\Pi_j^{s_j}$ being partnered.

Definition 2.1 (Correctness) *We call a group key establishment protocol $\mathcal{P}$* correct*, if in the presence of a passive adversary $\mathcal{A}$—i. e., $\mathcal{A}$ must not use the* Send *oracle—the following holds: for all i, j with $\mathsf{sid}_i^{s_i} = \mathsf{sid}_j^{s_j}$, $\mathsf{pid}_i^{s_i} = \mathsf{pid}_j^{s_j}$ and $\mathsf{acc}_i^{s_i} = \mathsf{acc}_j^{s_j} = \mathsf{true}$, we have $\mathsf{sk}_i^{s_i} = \mathsf{sk}_j^{s_j} \neq$*NULL.

Definition 2.2 (Strong security) *We say a group key establishment protocol is* strongly secure*, if the advantage $\mathsf{Adv}_{\mathcal{A}}$ of a ppt adversary $\mathcal{A}$ in attacking protocol $\mathcal{P}$ is a negligible function in the security parameter, where the advantage is defined as*

$$\mathsf{Adv}_{\mathcal{A}} := 2 \cdot \mathsf{Succ} - 1.$$

Here Succ *is the probability that the adversary queries* Test *only on fresh instances and guesses correctly the bit b used by the* Test *oracle in a moment when all these instances are still fresh.*

3 Proposal of a secure protocol

3.1 Signatures supporting randomness reveals

Our proposal of a secure protocol will make use of a signature scheme to authenticate the participants. As our security model allows the adversary to access the random coins involved in the execution of the protocol, by means of the oracle RevealRand, we assume that this oracle also outputs the randomness used for signing (if any). Therefore we introduce a new security notion

Set up:

Let $\mathcal{F}$ be a collision-resistant pseudorandom function family and v a fixed value. Further, let $\mathcal{UH}$ be a family of universal hash functions ranging in $\{0,1\}^\ell$, with ℓ such that $\{0,1\}^\ell$ is super-polynomial in the security parameter. A function UH from $\mathcal{UH}$ and F from $\mathcal{F}$ are made public together with a value v in the domain of F.

A pair of keys (pk_i, sk_i) for the signature scheme $\mathcal{S}$ is generated for each U_i, which gets the secret key sk_i while pk_i is publicized.

Round 0:

Usage of 2-SAKE.

- For $i = 1, \ldots, n$ execute 2-SAKE(U_i, U_{i+1}). After that each user U_i holds two keys $\overrightarrow{K_i}$ and $\overleftarrow{K_i}$ shared with U_{i+1} and U_{i-1} respectively.
- Additionally, in the last round of the 2-SAKE, each U_i chooses a random nonce $r_i \in_R \{0,1\}^\ell$, computes a signature σ_i^0 of (U_i, r_i) and broadcasts $M_i^0 := (U_i, r_i, \sigma_i^0)$.

Round 1:

Computation. Each U_i :

- Checks the signatures σ_j^0; if something fails, aborts;
- Sets $\mathsf{sid}_i := \mathsf{pid}_i | r_1 | \ldots | r_n$;
- Computes $X_i := \overrightarrow{K_i} \oplus \overleftarrow{K_i}$;
- Computes a signature σ_i^1 of $(U_i, \mathsf{sid}_i, X_i)$.

Broadcast. Each U_i broadcasts $M_i^1 := (U_i, \mathsf{sid}_i, X_i, \sigma_i^1)$.

Key Computation.

Check. Each U_i checks all the signatures, equality of pid's, sid's and $X_1 \oplus \cdots \oplus X_n = 0$; if something fails, aborts.

Computation. Each U_i

- for $j = 1, \ldots, n$, computes $\overleftarrow{K_j}$ and sets $K_j := \overleftarrow{K_j}$;
- sets $K := (K_1, \ldots, K_n, \mathsf{sid}_i)$;
- accepts $\mathsf{sk}_i := F_{UH(K)}(v)$.

Figure 1: A Compiler for achieving group AKE with strong security

for signature schemes, which we call *existential unforgeability under adaptive chosen message and randomness reveal attacks* (EUF−AMRA), that captures the property of remaining secure even if the randomness used when signing is leaked. Essentially we strengthen the standard security definition for signature schemes, i.e. existential unforgeability under adaptive chosen message attacks (EUF−AMA), by giving the adversary access to a more powerful oracle, that also provides the randomness used when generating the signature. A formal definition is depicted in the full version of this paper.

Note that this security notion is trivially achieved by a EUF−AMA signature scheme which is either deterministic or such that the randomness is provided to the verifier as part of the signature. As pointed out in [14], where an extensive list can be found, the latter is the case for many existing schemes.

3.2 From 2-Party to group keeping strong security: a compiler

In this section we present a compiler, which applied to a strongly secure 2-party key exchange 2-SAKE yields a strongly secure group key exchange adding only one communication round. Our construction is in the standard model, thus if the 2-SAKE does not involve any random oracle, so will the resulting n-party protocol be.

Our construction is detailed in Figure 1, where the **Set up** phase can be realized by means of a public key infraestructure (PKI) —and should thus be assumed to involve a trusted entity. Note that we further assume that there might be *independent* authentication keys used for the 2-party and group setting, namely, the compiler will call for (freshly generated) signing keys for a dedicated signature scheme (which we will denote by $(vk_i, sigk_i)$) while we also make explicit each user may have generated a pair of long-term keys $(2pk_i, 2sk_i)$ for 2-SAKE.[2] The n participants are arranged in a logical cycle and indexed modulo n. The construction fulfills the strong security notion depicted in Section 2:

Theorem 3.1 *Assuming $\mathcal{S}$ is an EUF−AMRA signature scheme and* 2-SAKE *is strongly secure, the protocol from Figure 1 is correct and also strongly secure.*

A formal proof of this statement is included in the full version of the paper. However, it is interesting to mention the reason that makes the scheme resilient to randomness reveals. A query RevealRand(U_i, s_i) returns $(\overrightarrow{r_i}, \overleftarrow{r_i}, r_i, sigr_i^0, sigr_i^1)$ where $\overrightarrow{r_i}, \overleftarrow{r_i}$ are the random coins used in the two executions of the 2-SAKE, r_i is the random nonce used in Round 1 of the compiler, and $sigr_i^j$, for $j = 0, 1$ are the nonces involved in the two signatures enforced by the compiler. Now our construction remains secure despite the RevealRandcalls, as it is easy to argue that this oracle is not useful for the adversary. Indeed, RevealRand(U_i) returns:

[2]This statement is quite general; note that these might not even be signing keys (as it would happen if 2-SAKE is the NAXOS scheme).

a) the randomness used by U_i in the 2-SAKE protocol, which is of no use for the adversary due to the strong security of 2-SAKE;

b) the signing nonces $sigr_i^0, sigr_i^1$, which will also be useless if the signature scheme is secure in the sense of EUF−AMRA;

c) the nonce r_i, which is anyway public, as it is broadcast in Round 1.

Acknowledgements

M.I. González Vasco and Ángel L. Pérez del Pozo are partially supported by research project MTM2013-41426-R, and A. Suárez Corona is supported by MTM2013-45588-C3-1-P, both funded by the Spanish MINECO.

References

[1] Michel Abdalla, Jens-Matthias Bohli, Maria Isabel Gonzalez Vasco, and Rainer Steinwandt. (password) authenticated key establishment: from 2-party to group. In *Theory of Cryptography*, pages 499–514. Springer, 2007.

[2] Michel Abdalla, Pierre-Alain Fouque, and David Pointcheval. Password-based authenticated key exchange in the three-party setting. In *Public Key Cryptography-PKC 2005*, pages 65–84. Springer, 2005.

[3] Florian Bergsma, Tibor Jager, and Jörg Schwenk. One-round key exchange with strong security: An efficient and generic construction in the standard model. In *Public-Key Cryptography–PKC 2015*, pages 477–494. Springer, 2015.

[4] Jens-Matthias Bohli, Maria Isabel Gonzalez Vasco, and Rainer Steinwandt. Secure group key establishment revisited. *Int. J. Inf. Sec.*, 6(4):243–254, 2007.

[5] Timo Brecher, Emmanuel Bresson, and Mark Manulis. Fully robust tree-diffie-hellman group key exchange. In *Cryptology and Network Security*, pages 478–497. Springer, 2009.

[6] Emmanuel Bresson and Mark Manulis. Securing group key exchange against strong corruptions. In *Proceedings of the 2008 ACM symposium on Information, computer and communications security*, pages 249–260. ACM, 2008.

[7] Cheng Chen, Yanfei Guo, and Rui Zhang. Group key exchange resilient to leakage of ephemeral secret keys with strong contributiveness. In *Public Key Infrastructures, Services and Applications*, pages 17–36. Springer, 2012.

[8] Cas JF Cremers. Session-state reveal is stronger than ephemeral key reveal: Attacking the naxos authenticated key exchange protocol. In *Applied Cryptography and Network Security*, pages 20–33. Springer, 2009.

[9] Atsushi Fujioka, Mark Manulis, Koutarou Suzuki, and Berkant Ustaoğlu. Sufficient condition for ephemeral key-leakage resilient tripartite key exchange. In *Information Security and Privacy*, pages 15–28. Springer, 2012.

[10] M Choudary Gorantla, Colin Boyd, Juan Manuel González Nieto, and Mark Manulis. Generic one round group key exchange in the standard model. In *Information, Security and Cryptology–ICISC 2009*, pages 1–15. Springer, 2009.

[11] M Choudary Gorantla, Colin Boyd, Juan Manuel González Nieto, and Mark Manulis. Modeling key compromise impersonation attacks on group key exchange protocols. *ACM Transactions on Information and System Security (TISSEC)*, 14(4):28, 2011.

[12] Brian LaMacchia, Kristin Lauter, and Anton Mityagin. Stronger security of authenticated key exchange. In *Provable Security*, pages 1–16. Springer, 2007.

[13] Mark Manulis, Koutarou Suzuki, and Berkant Ustaoglu. Modeling leakage of ephemeral secrets in tripartite/group key exchange. *IEICE Transactions*, 96-A(1):101–110, 2013.

[14] Sven Schäge. Strong security from probabilistic signature schemes. In *Public Key Cryptography–PKC 2012*, pages 84–101. Springer, 2012.

[15] Jianjie Zhao, Dawu Gu, and M Choudary Gorantla. Stronger security model of group key agreement. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 435–440. ACM, 2011.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Natural Convection MHD Stokes Flow in a Square Cavity

M. Gürbüz[1] and M. Tezer-Sezgin[1]

[1] *Department of Mathematics, Middle East Technical University*

emails: e160691@metu.edu.tr, munt@metu.edu.tr

Abstract

This study considers the steady natural convection slow flow of a viscous, incompressible and electrically conducting fluid in a square cavity under the effect of a uniform horizontally applied magnetic field. The cavity is heated from the left and the right wall is kept cooled while the other walls are adiabatic. The governing equations are obtained from the Navier-Stokes equations of fluid dynamics including buoyancy and Lorentz force terms and the energy equation including Joule heating and viscous dissipation terms. These equations are solved iteratively in terms of velocity components, stream function, vorticity, temperature and pressure by using radial basis function (RBF) approximation. Particular solution which is approximated by RBF to satisfy both differential equation and boundary conditions becomes the solution of the differential equation itself. For obtaining unknown vorticity boundary conditions, stream function equation is discretized using finite difference scheme which includes also interior stream function values. Pressure boundary conditions are taken as zero normal derivatives. The numerical results are obtained for Hartmann number (M) values in the range 0-50 and Grashof number (Gr) is taken 10 and 100 for fixed Reynolds number (Re) values of 0.6 . This way MHD Stokes flow is approached ($Re << 1$). As Gr increases, fluid moves to the right cold wall and isotherms concentrate trough the right boundary forming boundary layer. Dissipation of the viscous fluid alters the behaviour of the flow and temperature only when Gr is high. An increase in the magnetic field intensity retards the effect of buoyancy force. The solution is obtained in a considerably low computational cost trough the use of RBF approximations.

Key words: MHD Stokes flow, natural convection, viscous dissipation, RBF

1 Introduction

The natural convection flow under the influence of a magnetic field has many engineering applications such as the cooling of electronic systems, chemical processing equipment,

nuclear reactors, crystal growth and solar technology. In recent years, the effect of heat transfer on MHD flow is investigated by many researchers. Colaço et al. [1] have applied RBF approximation to solve incompressible MHD convection flow in terms of stream function and temperature neglecting Joule heating and viscous dissipation. MHD natural convection flow in a horizontal shallow cavity problem has been studied by Benos et al. [2]. Analytical solution was given with the method of matched asymptotic expansions. Kishore et al. [3] have showed the effect of viscous dissipation on MHD free convection flow past an exponentially accelerated vertical plate.

Highly viscous flow ($Re << 1$) in slow motion is called Stokes flow. Young et al. [4] have applied multiquadratic meshfree methods to solve 2D and 3D Stokes flow problem in a lid-driven and a circular cavity.

In the present study, we solve 2D steady, natural convection Stokes flow in a square cavity under a uniform magnetic field. The effects of both applied magnetic field and buoyancy force on Stokes flow are investigated in terms of all flow variables. Navier-Stokes and energy equations including Joule heating and viscous dissipation terms are solved iteratively by using RBF approximation which is easy to implement with small computational cost.

2 Mathematical formulation

MHD convection flow is considered in a square cavity when the temperature of the fluid is subjected to variations in the channel due to the wall temperature differences. A uniform magnetic field in the $x-$direction is applied. The continuity equation and the equations of motion for a natural convective MHD slow flow include buoyancy and Lorentz forces, and the energy equation includes Joule heating and viscous dissipation terms for an incompressible dissipative viscous fluid [6].
These are in $2D$, the contunity equation

$$\nabla \cdot \mathbf{u} = 0 \quad , \tag{1}$$

The momentum equation

$$\rho(\mathbf{u} \cdot \nabla)\mathbf{u} = -\nabla p + \rho\nu\nabla^2\mathbf{u} + \mathbf{J} \times \mu\mathbf{H} + \mathbf{g}\rho\beta(T - T_{cold}) \tag{2}$$

and the temperature equation

$$\rho c_p((\mathbf{u} \cdot \nabla)T) = \nabla \cdot (\lambda\nabla T) + \frac{1}{\sigma}\mathbf{J}^2 + \rho\nu\Phi + \rho Q \tag{3}$$

where $\mathbf{u} = (u, v)$, p, $\mathbf{H} = (H_x, H_y)$ and T are the velocity, pressure, the magnetic field and the temperature of the fluid, respectively. $\mathbf{g}\rho\beta(T - T_{cold})$ denotes the buoyancy force and Φ is the viscous dissipation function given by

$$\Phi = 2\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2.$$

Here ν, σ, μ, ρ, c_p, β and λ are kinematic viscosity, electrical conductivity, magnetic permeability, the density, specific heat, thermal expansion coefficient and thermal conductivity of the fluid, respectively. $\frac{1}{\sigma}\mathbf{J}^2$ is the Joule dissipation. Q represents other sources of volumetric energy release like nuclear radiation or chemical reactions. Physically, viscous dissipation is a transformation of kinetic energy into internal energy.

Now, we introduce the scaled transformations

$$\mathbf{x} \to \mathbf{x}L, \quad \mathbf{u} \to \mathbf{u}U_0, \quad \mathbf{H} \to \mathbf{H}H_0 \tag{4}$$

$$p \to p\rho\nu U_0/L, \quad Q \to Q_0 Q, \quad T - T_{cold} \to T(T_{hot} - T_{cold}) \tag{5}$$

and substitute into equations (1)-(3) to obtain the non-dimensional MHD convection equations as

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{6}$$

$$Re(u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}) = -\frac{\partial p}{\partial x} + \nabla^2 u \tag{7}$$

$$Re(u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y}) = -\frac{\partial p}{\partial y} + \nabla^2 v - M^2 v + \frac{Gr}{Re}T \tag{8}$$

$$\begin{aligned} Re(u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}) =& \frac{1}{Pr}\nabla^2 T + M^2 Ec v^2 \\ &+Ec\left(2\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2\right) + ReQ \end{aligned} \tag{9}$$

where $Re = LU_0/\nu$, $M = L\mu H_0\sqrt{\sigma/\rho\nu}$, $Ec = \dfrac{{U^2}_0}{c_p(T_{hot} - T_{cold})}$, $Gr = \dfrac{g\beta(T_{hot} - T_{cold})L^3}{\nu^2}$ and $Pr = \dfrac{\rho c_p \nu}{\lambda}$ are the Reynolds number, the Hartmann number, Eckert number, Grashof number and Prandtl number, respectively. Gr is the ratio of buoyancy force to viscous force. In our study, we consider air flow in cavities whose Pr is approximately equal to 0.71 and neglect the volumetric energy ($Q = 0$). If viscous dissipation is neglected, Eckert number is taken as $Ec = 0$.

Introducing 2D stream function ψ to satisfy the continuity equation as $u = \dfrac{\partial \psi}{\partial y}$, $v = -\dfrac{\partial \psi}{\partial x}$, and the vorticity as $\omega = \dfrac{\partial v}{\partial x} - \dfrac{\partial u}{\partial y}$, we obtain stream function equation $\nabla^2 \psi = -\omega$.

 ISBN: 978-84-608-6082-2

Differentiation of vorticity definition with respect to $x-$ and $y-$ gives velocity-vorticity equations as

$$\nabla^2 u = -\frac{\partial \omega}{\partial y} \quad , \qquad \nabla^2 v = \frac{\partial \omega}{\partial x} \quad . \tag{10}$$

Pressure Poisson's equation is obtained by differentiating equation (7) and equation (8) with respect to $x-$ and $y-$, respectively, and adding them up,

$$\nabla^2 p = -2Re(\frac{\partial v}{\partial x}\frac{\partial u}{\partial y} - \frac{\partial u}{\partial x}\frac{\partial v}{\partial y}) - M^2\frac{\partial v}{\partial y} + \frac{Gr}{Re}\frac{\partial T}{\partial y} \quad . \tag{11}$$

Similarly, vorticity Poisson's equation can be derived from cross-differentiation of equations (7) and (8) as

$$\nabla^2 \omega = Re(u\frac{\partial \omega}{\partial x} + v\frac{\partial \omega}{\partial y}) + M^2\frac{\partial v}{\partial y} - \frac{Gr}{Re}\frac{\partial T}{\partial x} \quad . \tag{12}$$

Thus, the $2D$ MHD convection flow is represented with Poisson's type equations in all the problem variables as velocity components, stream function, vorticity, pressure and temperature

$$\nabla^2 u = -\frac{\partial \omega}{\partial y} \quad , \qquad \nabla^2 v = \frac{\partial \omega}{\partial x} \tag{13}$$

$$\nabla^2 \psi = -\omega \tag{14}$$

$$\begin{aligned} \nabla^2 T =& PrRe(u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}) - M^2 EcPrv^2 \\ & -EcPr\left(2\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2\right) \end{aligned} \tag{15}$$

$$\nabla^2 \omega = Re(u\frac{\partial \omega}{\partial x} + v\frac{\partial \omega}{\partial y}) + M^2\frac{\partial v}{\partial y} - \frac{Gr}{Re}\frac{\partial T}{\partial x} \tag{16}$$

$$\nabla^2 p = -2Re(\frac{\partial v}{\partial x}\frac{\partial u}{\partial y} - \frac{\partial u}{\partial x}\frac{\partial v}{\partial y}) - M^2\frac{\partial v}{\partial y} + \frac{Gr}{Re}\frac{\partial T}{\partial y} \quad . \tag{17}$$

3 Numerical method (RBF approximation)

In the radial basis function approximation [5], the inhomogeneity in a partial differential equation $Lu(x,y) = f(x,y)$ is approximated as

$$f(x,y) \simeq \sum_{j=1}^{n} a_j \varphi_j(r), \quad (x,y) \in \Omega \tag{18}$$

and boundary condition is given as $Bu(x,y) = g(x,y)$ where L and B are linear operators, f and g are known functions in a domain Ω with the boundary $\partial\Omega$. $r = \sqrt{(x-x_j)^2 + (y-y_j)^2}$ is the Euclidean distance and n is the number of unknown coefficients, $\varphi_j(r)$'s are the radial basis functions.

Then, we can write the approximate solution $\hat{u}$ as

$$\hat{u}(x,y) = \sum_{j=1}^{n} a_j \Psi_j(r) \tag{19}$$

where $\{\Psi_j\}$ is obtained by back substitution through the differential equation $L\Psi_j(r) = \varphi_j(r)$ and with those $\{\Psi_j\}$, $\hat{u}$ is forced to satisfy the boundary condition $Bu = g$ as

$$\sum_{j=1}^{n} a_j B\Psi_j(r) = g(x,y), \qquad (x,y) \in \partial\Omega \quad . \tag{20}$$

The coefficients a_j in the approximation (19) are determined by taking $N + L = n$ collocation points (x_i, y_i) on the boundary and interior of the domain from the two linear systems

$$\sum_{j=1}^{n} a_j B\Psi_j(r_k) = g(x_k,y_k), \quad 1 \le k \le N \quad \text{and} \quad \sum_{j=1}^{n} a_j \varphi_j(r_l) = f(x_l,y_l), \quad 1+N \le l \le n \tag{21}$$

which are combined to give one linear system $[A]\{a\} = \{b\}$ for the solution vector $\{a\} = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}^T$. The coefficient matrix and the right hand side vector are given as

$$[A] = \begin{bmatrix} B\Psi_1(r_1) & B\Psi_2(r_1) & \cdots & B\Psi_n(r_1) \\ \vdots & \vdots & \ddots & \vdots \\ B\Psi_1(r_N) & B\Psi_2(r_N) & \cdots & B\Psi_n(r_N) \\ \varphi_1(r_{N+1}) & \varphi_2(r_{N+1}) & \cdots & \varphi_n(r_{N+1}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(r_n) & \varphi_2(r_n) & \cdots & \varphi_n(r_n) \end{bmatrix}_{n\times n} , \{b\} = \begin{bmatrix} g(x_1,y_1) \\ \vdots \\ g(x_N,y_N) \\ f(x_{N+1},y_{N+1}) \\ \vdots \\ f(x_n,y_n) \end{bmatrix}_{n\times 1} .$$

Solution of this system using Gaussian elimination gives the coefficients a_j's, $1 \le j \le n$, and $\hat{u}(x,y) = \sum_{j=1}^{n} a_j \Psi_j(r)$.

In this study, all the equations in (13)-(17) for MHD convection flow are considered as Poisson's type and solved iteratively approximating the right hand sides using RBF approximation. We take RBF as $\varphi(r) = 1 + r$, and so $\Psi(r) = \dfrac{r^2}{4} + \dfrac{r^3}{9}$. First, the equations

 ISBN: 978-84-608-6082-2

of velocity component (13) are solved with an initial estimate for vorticity. Also, stream function equation (14) is solved similarly. Then, with the use of the new values of velocity components and initial estimate for temperature, we solve the energy equation (15). Vorticity equation (16) is solved similarly. In each iteration, the required space derivatives of unknowns u, v, ω and T are obtained by using the coordinate matrix F as

$$\frac{\partial D}{\partial x} = \frac{\partial F}{\partial x} F^{-1} D, \qquad \frac{\partial D}{\partial y} = \frac{\partial F}{\partial y} F^{-1} D$$

where D denotes u, v, ω and T. The iteration continues until a preassigned tolerance (ϵ) is reached between two successive iterations. Then, we solve pressure equation (17) by using converged values of velocity components and temperature. Stream function equation is discretized using finite difference scheme which includes also interior values to obtain vorticity boundary conditions, whereas pressure boundary conditions are derived by using coordinate matrix for space derivatives and finite difference for pressure gradients. Normal derivative of pressure is taken as zero on the walls of the cavity.

4 Numerical results

The present iterative RBF approximation has been applied to natural convection flow of a dissipative viscous fluid in a square cavity under the effect of a uniform horizontally applied magnetic field. The problem configuration and the boundary values are shown in Figure 1.

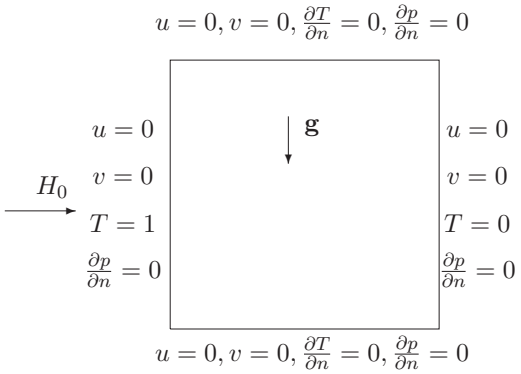


Figure 1: Square cavity and the boundary conditions

We take $N = 80$ (number of boundary points), $0 \leq M \leq 50$, and $Gr = 10, 100$ for fixed $Re = 0.6$ (as an approach to Stokes flow) and fixed Eckert number $Ec = 1$. The Prandtl number is taken as $Pr = 0.71$. The tolerance for stopping the iterations is 10^{-3}. Solution is depicted in terms of streamlines, vorticity, temperature and pressure contours in Figures 2-3.

Viscous dissipation effect on the MHD Stokes flow can be noticed for $Gr \geq 100$ in Figure 2. As Gr increases, dissipative viscous fluid ($Ec = 1$) moves to the cold right wall. When we look at the isotherms, heat spreads splitting through the center of the cavity with an increase in the magnitude, and then temperature value decreases and concentrates trough the right cold wall forming boundary layer. The flow and pressure behaviours are also altered. Flow circulates at the center but tends to move through the cold wall and pressure is symmetrically distributed in front of the walls when $Gr = 100$. Viscous dissipation retards the convection dominance on the temperature (bending of isotherms to be parallel to the adiabatic forms).

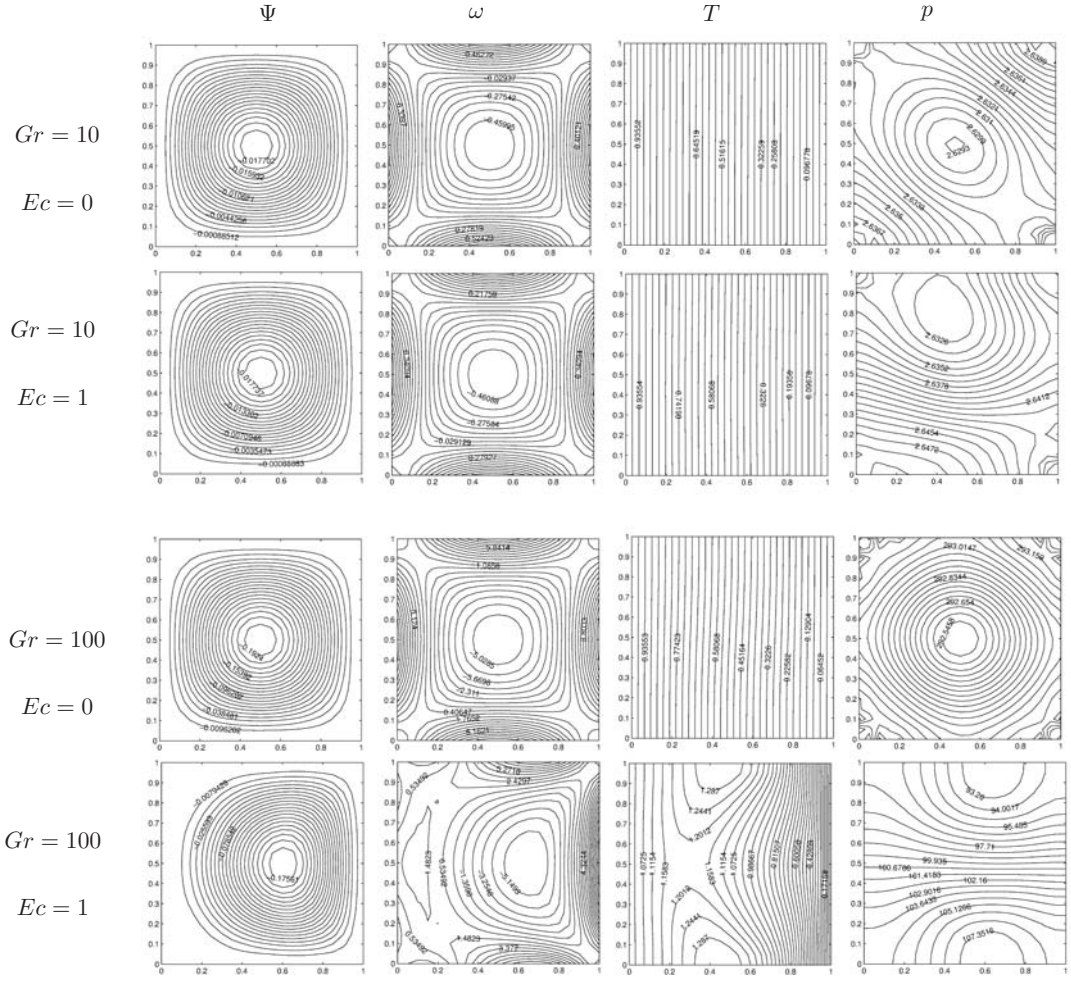


Figure 2: The effects of Gr on the flow, temperature and pressure for $M = 0$.

 ISBN: 978-84-608-6082-2

Figure 3 shows the effect of magnetic field on MHD natural convection Stokes flow when viscous dissipation is present. As M increases, center vortex of the streamlines converts to elliptic shape with a decrease in magnitude. Further an increase in M forces vorticity to be divided into two vortices in front of the bottom and top walls (adiabatic walls). Thin boundary layers are formed near the heated and cooled walls. Magnetic field retards the effect of the buoyancy force. When M reaches to the value of 50, the temperature comes back to the uniform distribution between the vertical walls. Moreover, an increase in M leads pressure profiles to form four symmetric vortices emanating from the corners.

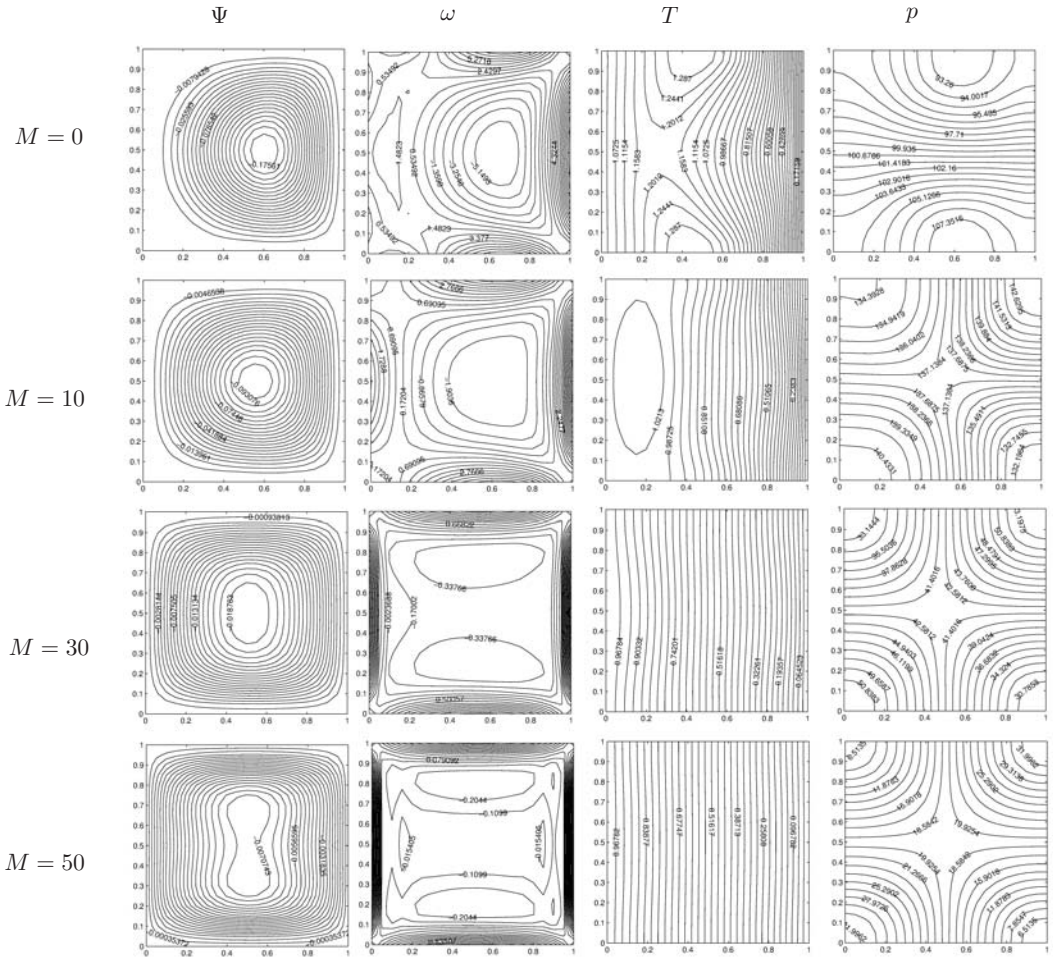


Figure 3: The effects of Hartmann number on the flow, temperature and pressure for $Gr = 100$.

 ISBN: 978-84-608-6082-2

5 Conclusion

The natural convection Stokes flow in a square cavity is solved under the effect of uniform horizontal magnetic field by using RBF approximation. The effects of buoyancy force and magnetic field on the behaviours of the flow, the temperature and pressure are investigated. The numerical results shows that increase in Grashof number results in an increase in the magnitude of the stream function and pressure. However, they decrease as Hartmann number increases. Magnetic field effect on isotherms is opposite to the effect of Grashof number. Viscous dissipation effect can be observed only when Gr reaches to 100.

Acknowledgement

This work has been partially supported by [the National Ph.D. Fellowship Programme of The Scientific and Technological Research Council of Turkey] under the Grant [2211].

References

[1] M. J. Colaço, G. S. Dulikravich, H. R. B. Orlande *Magnetohydrodynamic simulations using radial basis functions*, International Journal of Heat and Mass Transfer **52** (2009) 5932–5939.

[2] L. Th. Benos, S. C. Kakarantzas, I. E. Sarris, A. P. grecos, N. S. Vlachos, *Analytical and numerical study of MHD natural convection in a horizontal shallow cavity with heat generation*, International Journal of Heat and Mass Transfer **75** (2014) 19–30.

[3] P. M. Kishore, V. Rajesh, S. V. Verma *Effects of heat transfer and viscous dissipation on MHD free convection flow past an exponentially accelerated vertical plate with variable temperature*, Journal of Naval Architecture and Marine Engineering **7** (2010) 101–110.

[4] D. L. Young, S. C. Jane, C. Y. Lin, C. L. Chiu and K. C. Chen *Solutions of 2D and 3D Stokes laws using multiquadratics method*, Engineering Analysis with Boundary Elements **28** (2004) 1233–1243.

[5] C. S. Chen, C. M. Fan, P. H. Wen *The method of approximate particular solutions for solving certain partial differential equations*, Numerical Methods for Partial Differential Equations **28** (2012) 506–522.

[6] U. Müller and L. Bühler, *Magnetofluiddynamics in channels and containers*, Berlin, New York, 2001.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Solid State Materials With Transition-Metal Clusters and Fullerenes as Building Blocks

Lukas Hammerschmidt[1], Julia Schacht[1] and Nicola Gaston[2]

[1] *The MacDiarmid Institute for Advanced Materials and Nanotechnology, Victoria University of Wellington*

[2] *The MacDiarmid Institute for Advanced Materials and Nanotechnology, University of Auckland*

emails: lukas.hammerschmidt@vuw.ac.nz, julia.schacht@vuw.ac.nz, n.gaston@auckland.ac.nz

Abstract

Key words: superatoms, intercluster compounds, clusters, electronic structure, energetics, lattice structure, first-principles, DFT

1 Introduction

So-called superatoms are metal clusters, where the electronic structure is dominated by the number of valence electrons [1, 2]. Those valence electrons are mostly delocalized and follow a shell model [3, 4]. Upon shell closing, superatoms show an increased stability and atom-like behaviour, which allows them to act as bonding partners in molecules or solid state compounds [5, 6, 7]. Consequently, the capability to assemble such superatoms opens a broad field of new materials with a high potential for intriguing physical properties.

In 2013, Nuckolls et al. [7] reported three new materials (see Fig. 1), which are solid-state nanoscale-atom compounds based on CoSe-, CrTe- and NiTe-clusters in combination with fullerenes as bonding partners. The first two new compounds, $[Co_6Se_8(PEt_3)_6][C_{60}]_2$ and $[Cr_6Te_8(PEt_3)_6][C_{60}]_2$, showed an atom-like behaviour by forming a CdI_2 crystal structure. In CdI_2 cation layers are interchangeably either empty or fully occupied, which results in van-der-Waals bound neighbouring anion layers. Additionally, for both systems, experiments show indications of a charge transfer of one electron from the cluster to each of the two

 ISBN: 978-84-608-6082-2

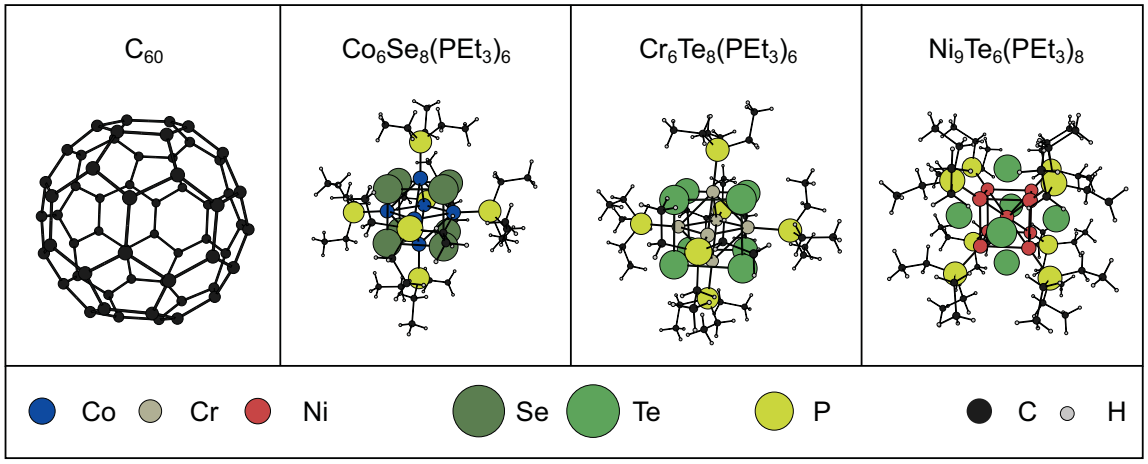


Figure 1: Schematic presentation of the $[Co_6Se_8(PEt_3)_6][C_{60}]_2$, $[Cr_6Te_8(PEt_3)_6]$, $[Ni_9Te_6(PEt_3)_8]$ cluster structures and the C_{60} fullerene. Atoms are not true to scale for reasons of clarity.

fullerenes. Both compounds show paramagnetic behaviour and the typical characteristics of small-gap semiconductors.

The third structure, $[Ni_9Te_6(PEt_3)_8][C_{60}]$, is quite different from the other two materials. It crystallizes in the cubic $Fm\bar{3}$ space group and for temperatures below 4 K it reveals ferromagnetic characteristics [8]. Additionally, in $[Ni_9Te_6(PEt_3)_8][C_{60}]$, fullerenes obtain even more charge compared to the CoSe- and CrTe-compounds.

Here, we have investigated all three of the aforementioned compounds and assess them for their superatom character in terms of their binding energies, electronic structure, and electronic transport by applying periodic DFT quantum chemistry methods.

2 Technical Details

We employed plane-wave DFT with the GW projector augmented wave (PAW) potentials as implemented in the VASP 5.3.5 program package [9, 10, 11, 12]. Full structure relaxations, electronic structures and related properties were obtained from spin-polarized LDA [13], PBE [14], PBEsol [15], PBE-D3 [16] and PBE-DTS [17] computations.

Charges were obtained by the Bader formalism. Transport properties were obtained semi-classically by Boltzmann transport theory and the first principles band structures. Projections onto spherical harmonics for the superatom assessment were performed as described earlier [18], with spheres of appropriate sizes for all three materials, respectively.

The cohesive energies are obtained as

$$E_{\mathrm{coh}} = E_{\mathrm{Compound}} - E_{\mathrm{Cluster}} - nE_{\mathrm{C_{60}}}, \tag{1}$$

the difference of the energy of the compound (E_{Compound}) and the energies of its building blocks, i.e. the ligand-protected cluster (E_{Cluster}) and the n fullerenes ($E_{\mathrm{C_{60}}}$) in the unit cell.

3 Results

Due to the size – small variations in large size unit cells may add up to large errors – for all three compounds relaxation of the unit cells is essential. All functionals perform reasonably well in describing the lattice parameters, within the limitations of the applied exchange-correlation functionals (see Fig. 2). It becomes apparent that dispersion corrections are important due to large contributions of van-der-Waals interactions. Thus, great improvements on the lattice structure are achieved for PBE-D3 and PBE-DTS. An outstanding and surprisingly excellent agreement compared to experiment is reached by the PBEsol functional for all compounds. As described in the introduction, van-der-Waals interactions are

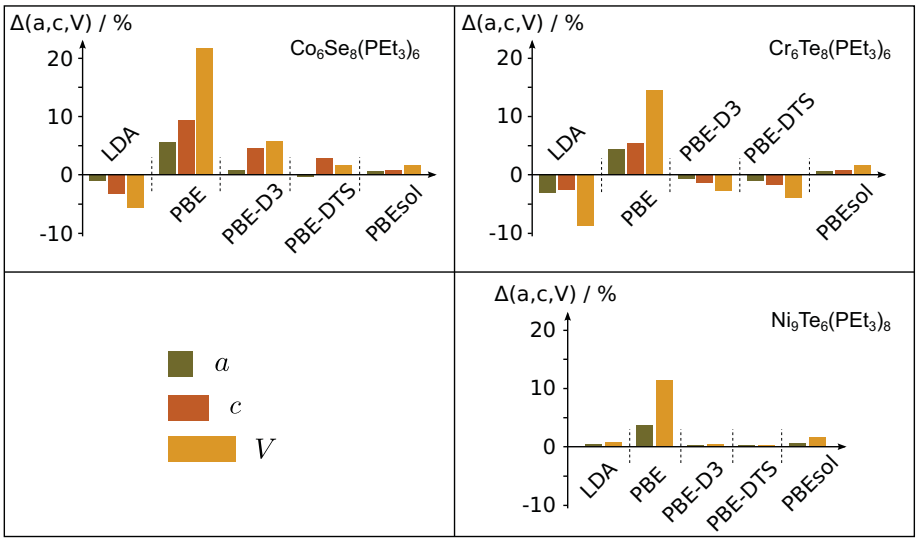


Figure 2: Relative deviations of computed lattice parameters in comparison to experiment [7] as obtained by various exchange correlation functionals within DFT and the dispersion corrected PBE versions of Grimme and Tkatchenko.

 ISBN: 978-84-608-6082-2

smaller in the NiTe-based compound. Consequently, in that case all functionals perform better.

Cohesive energies confirm the strong van-der-Waals contributions in the bonding. Here, we use the cohesive energy to analyze the bonding, rather than to compare to experiment. All applied functionals predict strongly deviating cohesive energies. Where standard functionals tend to (strongly) underestimate the binding, dispersion corrected functionals seem to overestimate the bonding. Better agreement between the various methods is achieved for the NiTe-compound, which is again due to less van-der-Waals contributions.

However, although there is an obvious strong contribution of dispersive interactions, we observe an additional charge transfer, similar to experiment. According to our charge analysis about one electron is transfered from the metal clusters to the fullerenes. This charge separation is in fact enabled by the metal core-protecting ligands. The charge separation leads to an additional stabilization of the compounds by electrostatic interactions.

Band structure computations reveal small-gap semiconductors for the CoSe- and CrTe-compounds in agreement with experimental observations. A strong spin-polarization within the NiTe-compound indicates a magnetic ground state which is indeed observed in experiment for temperatures below 4 K. Our computed semi-classical transport properties agree well with experiments allowing for a properties evaluation.

4 Conclusions

Our results show that the lattice structure of our presented superatom compounds relates strongly to actual atoms. Not only crystallize the large building blocks according to their smaller relatives but additionally a similar amount of charge is transfered between the "atoms". That the large building blocks form an actual solid, is reflected by the compounds' transport properties. The electronic structure, however, shows different characteristics than would be expected from a typical superatomic shell closing.

Acknowledgements

This work has been supported by the MacDiarmid Institute for Advanced Materials and Nanotechnology and the New Zealand eScience Infrastructure (NeSI).

References

[1] W. A. de Heer, *The physics of simple metal clusters: experimental aspects and simple models*, Rev. Mod. Phys. **65** (1993) 611–675.

[2] W. D. Knight, K. Clemenger, W. A. de Heer, W. A. Saunders, M. Y. Chou, and M. L. Cohen, *Electronic Shell Structure and Abundances of Sodium Clusters*, Phys. Rev. Lett. **52** (1984) 2141–2143.

[3] J. Pedersen, S. Bjornholm, K. Hansen, T. P. Martin, and H. D. Rasmussen, *Observation of quantum supershells in clusters of sodium atoms*, Nature **353** (1991) 733–735

[4] M. Walter, J. Akola, O. Lopez-Acevedo, P. D. Jadzinsky, G. Calero, C. J. Ackerson, R. L. Whetten, H. Groenbeck, and H. Haekkinen, *A unified view of ligand-protected gold clusters as superatom complexes*, Proc. Natl. Acad. Sci. **105** (2008) 9157–9162.

[5] D. E. Bergeron, A. W. Castleman Jr., T. Morisato, and S. N. Khanna, *Formation of $Al_{13}I^-$: Evidence for the Superhalogen Character of Al_{13}*, Science **304** (2004) 84–87.

[6] M. Schulz-Dobrick and M. Jansen, *Intercluster Compounds Consisting of Gold Clusters and Fullerides: $[Au_7(PPh_3)_7]C_{60}\cdot THF$ and $[Au_8(PPh_3)_8](C_{60})_2$*, Angew. Chem. Int. Ed. **47** (2008) 2256–2259.

[7] X. Roy, C.-H. Lee, A. C. Crowther, C. L. Schenck, T. Besara, R. A. Lalancette, T. Siegrist, P. W. Stephens, L. E. Brus, P. Kim, M. L. Steigerwald, and C. Nuckolls, *Nanoscale Atoms in Solid-State Chemistry*, Science **341** (2013) 157–160.

[8] Chul-Ho Lee, L. Liu, C. Bejger, A. Turkiewicz, T. Goko, C. J. Arguello, B. A. Frandsen, S. C. Cheung, T. Medina, T. J. S. Munsie, R. DÓrtenzio and G. M. Luke, T. Besara, R. A. Lalancette, T. Siegrist, P. W. Stephens, A. C. Crowther, L. E. Brus, Y. Matsuo, E. Nakamura, Yasutomo, J. Uemura, P. Kim, C. Nuckolls, M. L. Steigerwald, and X. Roy, *Ferromagnetic Ordering in Superatomic Solids*, J. Am. Chem. Soc. **136** (2014) 16926–16931.

[9] G. Kresse and J. Furthmüller, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Comput. Mat. Sci. **6** (1996) 15.

[10] G. Kresse and J. Furthmüller, *Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set*, Phys. Rev. B **54** (1996) 11169.

[11] G. Kresse and J. Hafner, *Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium*, Phys. Rev. B **49** (1994) 14251.

 ISBN: 978-84-608-6082-2

[12] G. Kresse and J. Hafner, *Ab initio molecular dynamics for liquid metals*, Phys. Rev. B **47** (1993) 558.

[13] J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B **23** (1981) 5048.

[14] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77** (1996) 3865.

[15] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, K. Burke, *Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces*, Phys. Rev. Lett. **100** (2008) 136406.

[16] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu*, J. Chem. Phys. **132** (2010) 154104.

[17] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data*, Phys. Rev. Lett. **102** (2009) 073005.

[18] D. Schebarchov and N. Gaston, *Electronic shell structure in Ga12 icosahedra and the relation to the bulk forms of gallium*, Phys. Chem. Chem. Phys. **14** (2012) 9912–9922.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Cloud implementation of the K-means algorithm for hyperspectral image analysis

Juan Mario Haut[1], Mercedes Paoletti[1], Javier Plaza[1] and Antonio Plaza[1]

[1] *Department of Technology of Computers and Communications, University of Extremadura, Escuela Politecnica, Avda. de la Universidad s/n*

emails: juanmariohaut@unex.es, mpaolett@alumnos.unex.es, jplaza@unex.es, aplaza@unex.es

Abstract

Remotely sensed hyperspectral imaging offers the possibility to collect hundreds of images, at different wavelength channels, for the same area on the surface of the Earth. Hyperspectral images are characterized by their large volume and dimensionality, which makes their processing and storage difficult. As a result, several techniques have been developed in previous years to perform hyperspectral image analysis on high performance computing architectures. However, the application of cloud computing techniques has not been as widespread. There are many potential advantages in exploiting cloud computing architectures for distributed hyperspectral image analysis. In this paper, we present a cloud implementation (developed using Apache Spark) of the popular K-means algorithm for unsupervised hyperspectral image clustering. The experimental results suggest that cloud architectures allow for the efficient distributed processing of large hyperspectral image data sets.

Key words: hyperspectral imaging, cloud computing, k-means clustering.

1 Introduction

Hyperspectral images comprise hundred of contiguous spectral bands, thus imposing significant requirements in terms of storage and data processing. These requirements have increased exponentially with the technological advances in satellite and airbone remote sensing, leading to the creation of high-dimensional hyperspectral data repositories [1].

The availability of new hyperspectral missions is now generating a continuous stream of multi/hyperspectral data, and this has introduced important challenges for scalable and

efficient processing of hyperspectral data in the context of different applications [1]. For instance, the NASA Jet Propulsion Laboratory's Airbone Visible/Infrared Imaging Spectrometer (AVIRIS) [2] has a data collection rate of 2.5 MB/s (nearly 9 GB/hour). A similar case is the space-borne Hyperion instrument [1], which collects almost 71.9 GB/hour (over 1.6 TB/day). Most of the satellite missions that will be soon in operation, such as the environmental mapping and analysis program (EnMAP)[1] present similar data collection ratios. Hyperspectral data repositories are becoming increasingly massive and often distributed among several geographic locations due to their volume, which makes hard to meet the storage and computational requirements of large-scale hyperspectral data processing applications without resorting to distributed computing facilities.

In recent years, cloud computing platforms have been adopted for remotely sensed data processing. The cloud is now a standard for distributed computing due to its advanced capabilities for internet-scale computing, service-oriented computing, and high-performance computing. The use of cloud computing for the analysis of large hyperspectral data repositories can be considered a natural solution and an evolution of previously developed techniques for other kinds of computing platforms [3]. However, there are few efforts in the recent literature oriented to the exploitation of cloud computing infrastructure for hyperspectral imaging techniques in general, and for unsupervised clustering algorithms in particular.

Clustering can be defined as a segmentation process in which pixels are assigned into a group that represents a specific land-cover class [4]. The main advantage of clustering is that there is no need for labeled samples which are common in supervised classification techniques [5]. In this regard, clustering offers an unsupervised alternative that has been widely used in various fields. However, clustering is also a very challenging task due to the large spectral variability and complex spatial structures present in hyperspectral images. A widely used family of clustering algorithms is represented by centroid-based clustering methods such as K-means [4], which assumes that similar pixels always form clusters in feature space. When applied to hyperspectral images, these methods can provide satisfactory results but are hampered by their large computational complexity.

In this paper we explore the possibility of using a distributed framework for massive hyperspectral image processing based on cloud computing. We use unsupervised clustering as a case study, focusing on the K-means algorithm to demonstrate the applicability of utilizing cloud computing technologies to efficiently perform distributed parallel processing of hyperspectral data and accelerate computations.

The remainder of the paper is organized as follows. Section II presents the distributed framework design that will be used in our implementation. Section III presents the K-means algorithm and its distributed implementation. Section IV experimentally assesses the proposed method in terms of both accuracy and computational performance. Finally, section IV concludes with some remarks and hints at future research lines.

[1] http://www.enmap.org/

 ISBN: 978-84-608-6082-2

2 Distributed framework design

In order to develop a distributed framework for implementing the K-means algorithm on cloud computing architectures, two main issues need to be addressed: 1) the distributed programming model and 2) the computing engine.

For distributed programming, we resort to the MapReduce model [3], taking full advantage of the high-performance capabilities provided by cloud computing architectures. In this model, a task is processed by two distributed operations: map and reduce. The datasets are organized as key/value pairs, and the map function processes a key/value pair to generate a set of intermediate pairs, dividing a task into several independent subtasks to be run in parallel. The reduce function is in charge of processing all intermediate values associated with the same intermediate key, then collecting all the subtask results to gather the result for the whole task.

Regarding the distributed computing engine, a first solution considered was Apache Hadoop[2] due to its reliability and scalability, as well as its completely open source nature. However, Apache Hadoop only supports simple one-pass computations and is generally not appropriate for iterative algorithms such as K-means. Apache Spark[3] is a newly developed computing engine for large-scale data processing on cloud computing architectures, which implements a fault-tolerant abstraction for in-memory cluster computing, and provides fast and general data processing on large distributed platforms. It not only supports simple one-pass computations, but can also be extended to the case of multi-pass, iterative algorithms.

With the aforementioned issues in mind, the design of our distributed parallel framework for hyperspectral data clustering using Apache Spark is graphically sketched in Fig. 1.

As shown by Fig. 1, the architecture has two main parts:

- The hardware zone: it contains the physical machines that support our virtual machines, which are created in the OpenStack[4] platform, a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed through a dashboard that gives administrators control while empowering their users to provision resources through a web interface.

- The platform zone: the Apache Spark framework in Fig. 1) is installed over a set of Ubuntu Linux virtual machines, created by OpenStack. Our cluster has various types of nodes. The K-means algorithm is designed using the MLib machine learning library [5], and the implementation is embedded into a joint Spark and OpenStack framework, as illustrated graphically in Fig. 2. When we launch an instance of K-means, the master node manages the resources of the cluster and the slaves (workers) perform

[2]http://hadoop.apache.org
[3]http://spark.apache.org/
[4]https://wiki.openstack.org/wiki/Main_Page
[5]https://spark.apache.org/docs/latest/mllib-guide.html

individual tasks on the data. The driver node divides the work into tasks, and the master node coordinates the allocation of tasks with the idea that all the task will be executed in the worker nodes by the executors, following the MapReduce model.

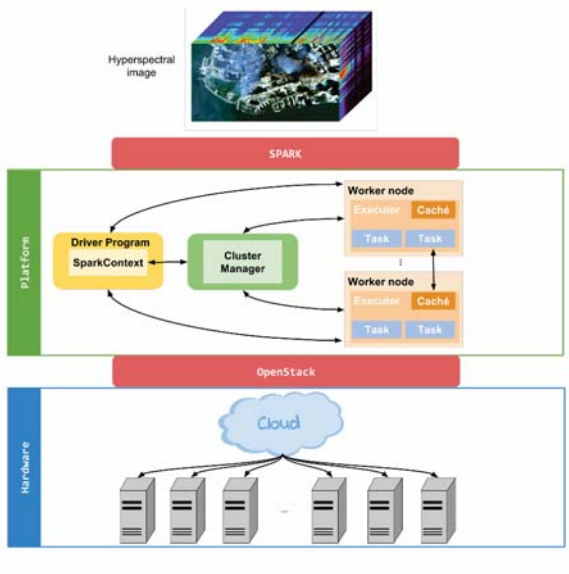


Figure 1: Integrated Apache Spark and OpenStack framework used for the implementation of K-means.

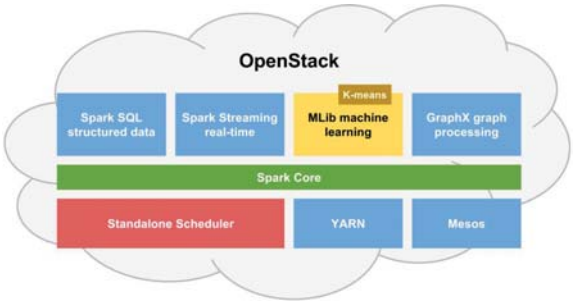


Figure 2: Description of the Apache Spark architecture used in our experiments.

3 The K-means clustering algorithm

K-means is one of the most widely used unsupervised algorithms to group data in a specified number of clusters. The procedure begins with a set of data or observations, $X = [x_1, x_2, ..., x_n]$, in $\mathbb{R}^d$ (so $x_i = [x_{i1}, x_{i2}, ..., x_{id}]$, with $i = 1, 2, ..., n$), that needs to be grouped into a number of *clusters* ($k <= n$). Iteratively, K-means calculates the centers of the k groups, optimizing the error of each group as follows:

$$\min \sum_{j=1}^{k} \sum_{i=1}^{n_k} \parallel x_i^j - C_j, \parallel^2 \tag{1}$$

where $\parallel x_i^j - C_j \parallel^2$ is the distance between a data point x_i^j of the cluster j (n_k is the number of observations within each cluster) and the cluster center C_j.

 ISBN: 978-84-608-6082-2

The K-means algorithm strongly depends on the choice of the initial centers (K-means can converge to a local minimum). So, a proper initialization will result in a final best solution. In order to obtain a set of good initial cluster centers, several methods have been proposed. One of them is the K-means++ method [6]. This algorithm obtains a set of k initial centers which are generally very close to the final solution by the following five steps:

1. An initial point is chosen from the set of samples $X = [x_1, x_2, ..., x_n]$ by an uniform random variable. This point c_1, is the first center.

2. For each sample x_i, the distance, $D(x)$, between x_i and the center, C_1, is calculated.

3. Then, a new candidate to become center is randomly selected using the probability-weighted distribution $\frac{D(x)^2}{\sum_{i=0}^{k} D(x_i)^2}$.

4. Steps 2 and 3 are repeated until k initial centers have been selected.

5. Once the k initial centers have been chosen, we apply the standard K-means algorithm.

4 Parallel/distributed implementation of K-means

4.1 Distributed implementation on Apache Spark

Apache Spark implements a parallel version of the K-means++ method, called *k-means||* [7] using the MapReduce model of computation. K-means|| is very similar to K-means++ with the difference that, instead of choosing a single point after calculating the probability distribution of each of the points in the data set, several clusters are chosen at each iteration. At the end of the parallel algorithm $O(k \log n)$ points are obtained, which are clustered in k centers. This process is illustrated in Algorithm 1.

In order to use the K-means method in Apache Spark, the user must specify the following parameters: number of desired clusters, k; maximum number of iterations that the algorithm can be executed, $maxIterations$; and number of times to execute the K-means algorithm completely, $runs$[6]. The $initializationMode$ indicates the type of initialization (if completely random or with K-means|| method. Also, a set of initial centers can be introduced using the $initialModel$ option. Finally, the number of steps to be executed during the K-means++ phase, $initializationSteps$, and a pre-defined error threshold for convergence, ϵ, should also be specified.

[6]Since the K-means may not find the optimal overall solution, it is recommended to run it several times to converge to a better final solution.

Algorithm 1 K-means|| algorithm

procedure K-means||(k, l) ▷ $k \rightarrow$ number of clusters, $l \rightarrow \Theta(K)$
  $C \leftarrow$ uniform_rand(X)
  $\psi \leftarrow \phi_X(C)$
  for $O(\log \psi)$ **do**
    $C' \leftarrow x \in X$ with $p(x) = \frac{l \cdot D(x)^2}{\phi_X(C)}$ ▷ Probability-weighted distribution
    $C \leftarrow C \bigcup C'$
  end for
  For $x \in C$, set w_x to be the number of points in X closer to x than any other point in C
  Reclusted (C, k)
end procedure

4.2 Paralllel implementation on Scikit-Learn

Scikit is a Python Library for machine learning. This library contains its own k-means++ version, as an option of the *kmeans* class into the *sklearn.cluster* package, allowing the parallel execution of K-means method with the parameter *n_jobs*. The operation is quite similar to Spark, since the user must indicate the number of clusters, *n_clusters*; the number of iterations of one complete execution of the K-means algorithm, *max_iter*; and the number of times the k-means will run with different centers, *n_init*[7]. Much like Apache Spark, Scikit has the option to initialize the first set of centroids randomly, using k-means++ or through an array of data, called *init*. Finally, we must indicate the *tol* or relative increment in the results before declaring convergence, and the number of CPUs that will be used during the execution, *n_jobs*.

5 Experiments

5.1 Hyperspectral datasets

The images used in our experiment were collected by the Airborne Visible-Infrared Imaging Spectrometer (AVIRIS) [2]. The well-known Indian Pines dataset was adquired in 1992 over an agricultural site composed of agricultural fields with regular geometry and with a multiple crops:

1. The first scene has a size of 145x145 pixels, and it was collected over a mixed forest and agriculture area. It has 220 spectral bands in the range from 400 to 2500 nm, with spectral resolution of 10 nm, moderate spatial resolution of 20 nm, and 16 bits

[7] After iterating, the algorithm takes only the best solution reached.

radiometric resolution. After an initial analysis, 8 bands have been removed due to noise, ending up with a total of 212 bands. About half of the pixels in the image (10366 of 21025) contain ground-truth information, which comes in the form of a single label assignment having a total of 16 ground-truth classes.

2. The second scene has a much larger size of 2678x614 pixels. It was collected over the same area, but spanning a much larger extent. It contains 220 spectral bands in the range from 400 to 2500 nm, with spectral resolution of 10 nm, moderate spatial resolution of 20 nm and 16 bits of radiometric resolution. The percentage of pixels with ground truth information is 20.33% (334245 out of 1644292 pixels) and the total number of classes is 58.

5.2 Experimental Configuration

The distributed environment in which we have tested our implementation is the one described in section 2. As mentioned in that section, the cloud computing platform used for experimental evaluation OpenStack. This software is a collection of Open Source technologies that provide a scalable deployment of a cloud computing environment. Our environment is composed of Intel(R) Xeon(R) CPUs E5430 @ 2.66GHz (8 cores), 16 GB RAM, Shared storage, NetApp FAS3140.

The virtual nodes that we use over the hardware specified have two virtual CPUs, 4GB of RAM and 40 GB hard disk each. In addition, we have developed a parallel version of the algorithm for comparative purposes. This version has been implemented on a paltform with Intel(R) Core(TM) i7-4790 CPUs @ 3.60GHz (8 cores), 16 GB RAM, SanDisk SDSSDA240G.

In our experiments, we used Java 1.8.0_92-b14, Ubuntu 14.04 x64 LTS as operating system, Python 2.7.10, Scikit Learn 0.14.1 version and the newest (under-development) Apache Spark 2.11.

5.3 Description of experiments

5.3.1 Single vs multiple cores

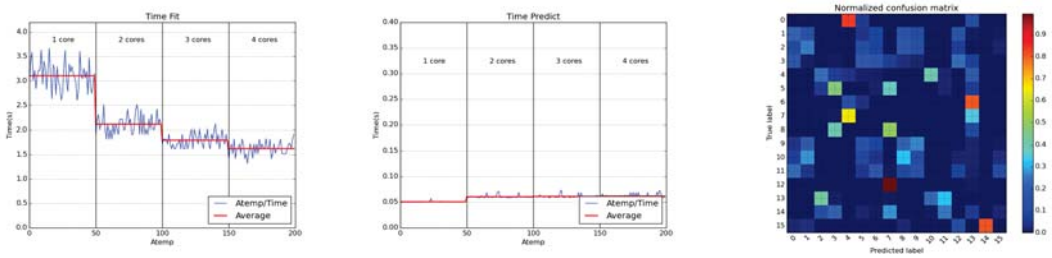
In a first experiment, we considered the small Indian Pines image and launched 50 executions of the algorithm changing the number of cores between 1 and 4. In this test, we set the tolerance threshold to 1e-15, we performed 10 iterations with different centroid seeds, and used k-means++ to select initial cluster centers. The obtained results are summarized in Table 1, which reports on the learning fit, the prediction accuracy, the clustering accuracy and the reliability (average and standard deviation). As shown in this table, the values remain constant as we increase the number of cores except the trainning of the model. The fit time is reduced until 3 cores are used (see Fig. 3a). At that point, there is no difference

to add more workers because the algorithm no longer is able to parallelize more information and therefore the speedup remains constant.

To validate the results obtained from K-means++, we used a confusion matrix [8] which is graphically represented in Fig. 3c. The confusion matrix indicates the agreement between the ground-truth classes and the clusters identified in the process, which appears to indicate a good fit between each ground-truth class and at least one of the identified clusters. Finally, the clusters obtained by the K-means++ algorithm for the small Indian Pines image are shown in Fig. 4. The figure shows the clusters with the background of the image removed (i.e., those pixels that do not have associated ground-truth) and also without removing the background.

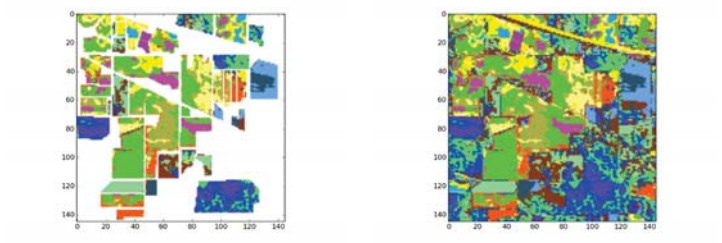
Cores	AVG Fit	Std Fit	AVG Predict	Std Predict	AVG Accuracy	Std Accuracy	AVG Reliability	Std Reliability
1	3.1055	0.2826	0.0508	0.0010	0.5446	0.0019	0.4770	0.0058
2	2.1135	0.1846	0.0606	0.0028	0.5456	0.0029	0.4772	0.0067
3	1.7955	0.1387	0.0612	0.0028	0.5452	0.0022	0.4770	0.0059
4	1.6213	0.1282	0.0620	0.0037	0.5453	0.0017	0.4779	0.0054

Table 1: Summary of the execution of K-means in multiple cores with the small Indian Pines image.



(a) Time to training (b) Time to clustering (c) Confusion matrix

Figure 3: Graphical representation of the time to train, time to clustering and the confusion matrix obtained after applying the K-means algorithm to the small Indian Pines image



(a) Without background (b) With background

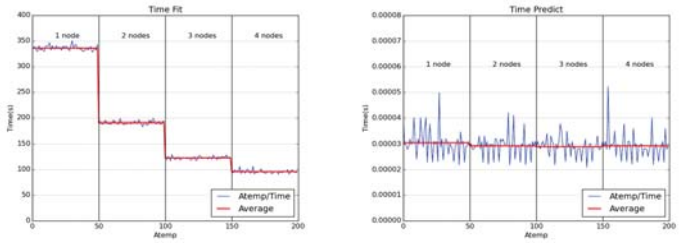
Figure 4: Clustering results obtained with the small Indian Pines image

5.3.2 Single vs multiple nodes

In this experiment, we launch 50 executions with the large Indian Pines image changing the number of slave nodes to 1, 2, 3 and 4. We select a tolerance threshold of 1e-15 and 10 run with different centroid seeds and using the k-means|| algorithm. Here, we use the large Indian Pines data set. The results obtained are shown in Table 2. In the table, we can observe how the values remain constant as we increase the number of nodes (except the trainning model). The fit time is reduced exponentialy (see Fig. 5a). Moreover, to validate the results obtained from K-Means, we show the confusion matrix for the large Indian Pines image in Fig. 6a. The final clusters obtained by the algorithm are shown in Fig. 7.

Nodes	AVG Fit	Std Fit	AVG Predict	Std Predict	AVG Accuracy	Std Accuracy	AVG Reliability	Std Reliability
1	336.0076	4.5684	0.0000305	0.00000522	0.4118	0.0132	0.2547	0.0043
2	191.1449	3.0629	0.0000293	0.00000415	0.4107	0.0153	0.2543	0.0046
3	122.8018	2.8471	0.0000289	0.00000335	0.4149	0.0114	0.2542	0.0050
4	95.9846	3.2338	0.0000293	0.00000502	0.4127	0.0120	0.2536	0.0048

Table 2: Summary of the execution of K-means in multiple nodes with the large Indian Pines image.



(a) Time to training (b) Time to clustering

Figure 5: Confusion matrix obtained after applying the K-means algorithm to the large Indian Pines image.

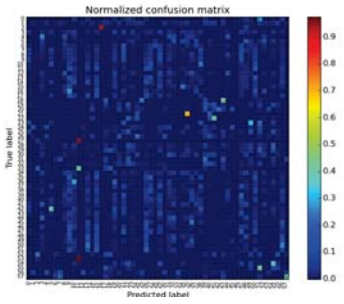


Figure 6: Graphical representation of the time to train and time to clustering after applying the K-means algorithm to the large Indian Pines image.

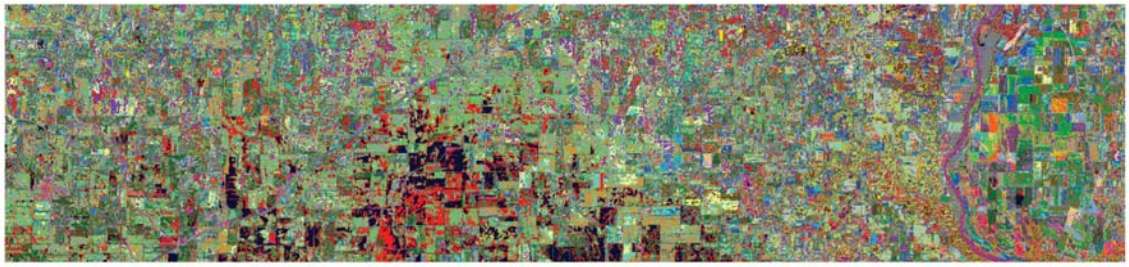


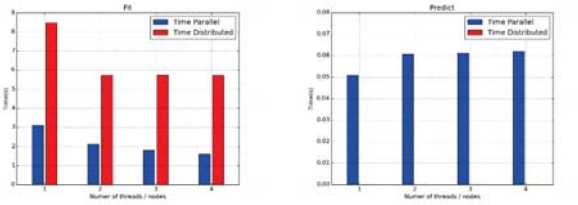
Figure 7: Clustering results obtained with the large Indian Pines image.

5.4 Parallel vs distributed

Finally we compare the parallel and distributed implementations of K-means. Table 3 shows the timing average (in seconds) measured after the parallel/distributed execution of K-means using both images and multiple cores/nodes. Figs. 8 and 9 respectively provide a graphical representation of the times for fitting and predicting for the parallel and distributed version with both images. Finally, we show in Fig. 10 a graphical representation of speedup evolution in all cases.

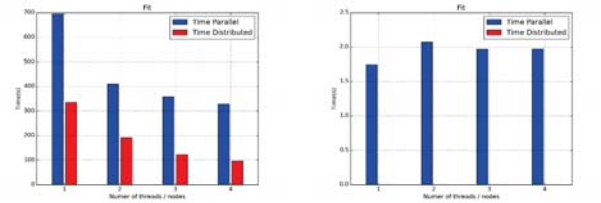
Image used	small Indian Pines image				large Indian Pines image			
Cores or nodes	**Par Fit**	**Par Predict**	**Dist Fit**	**Dist Predict**	**Par Fit**	**Par Predict**	**Dist Fit**	**Dist Predict**
1	3.1055	0.0508	8.4701	0.0000279	694.8973	1.7438	336.0076	0.0000305
2	2.1135	0.0606	5.7111	0.0000278	409.5856	2.0710	191.1449	0.0000293
3	1.7955	0.0612	5.7370	0.0000289	357.4873	1.9692	122.8018	0.0000289
4	1.6213	0.0620	5.7213	0.0000287	327.6066	1.9747	95.9846	0.0000293

Table 3: Timing average (in seconds) measured after the parallel/distributed execution of K-means using both images and multiple cores/nodes (Par refers to the parallel version and Dist refers to the distributed version).



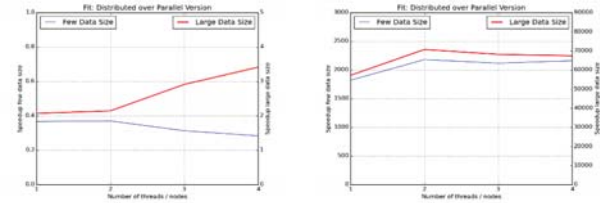
(a) Fitting (b) Predicting

Figure 8: Graphical representation of the times for fitting and predicting with the small Indian Pines image for the parallel and distributed implementation of K-means.



(a) Fitting (b) Predicting

Figure 9: Graphical representation of the times for fitting and predicting with the large Indian Pines image for the parallel and distributed implementation of K-means.



(a) Speedup (fitting) (b) Speedup (predicting)

Figure 10: Graphical representation of the speedup for fitting and predicting with the Indian Pines images for the parallel and distributed implementation of K-means.

6 Conclusions and future lines

In this paper, we have discussed the possibility of exploiting cloud computing architectures for hyperspectral image processing. As a case study, we have presented a cloud computing implementation of the K-means algorithm on Spark platform. Clustering has the advantage that it can be performed in unsupervised fashion. Our experimental results show the effectiveness of the proposed distributed implementation, not only in terms of clustering accuracy but also in terms of computational performance. As future work, we will implement other algorithms for hyperspectral data processing.

Acknowledgements

This work has been supported by Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005). This work was supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional

Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain.

References

[1] Antonio Plaza, Javier Plaza, Abel Paz, and Sergio Sanchez, "Parallel Hyperspectral Image and Signal Processing," *IEEE Signal Processing Magazine*, vol. 28, pp. 196–218, 2011.

[2] Robert O. Green, Michael L. Eastwood, Charles M. Sarture, Thomas G. Chrien, Michael Aronsson, Bruce J. Chippendale, Jessica A. Faust, Betina E. Pavri, Christopher J. Chovit, Manuel Solis, Martin R. Olah, and Orlesa Williams, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, pp. 227–248, 1998.

[3] Zebin Wu, Yonglong Li, Antonio Plaza, Jun Li, Fu Xiao, and Zhihui Wei, "Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–9, 2016.

[4] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm:Analysis and Implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881 – 892, 2002.

[5] Antonio Plaza, Javier Plaza, Gabriel Martin, and Sergio Sanchez, "Hyperspectral Data Processing Algorithms," in *Hyperspectral Remote Sensing of Vegetation*, Alfredo Huete Prasad S. Thenkabail, John G. Lyon, Ed., chapter 5, pp. 121–137. Taylor and Francis, Abingdon, United Kingdom, 2011.

[6] David Arthur and Sergei Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in *Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM, Ed., New Orleans, Louisiana, 2007, pp. 1027–1035, Society for Industrial and Applied Mathematics.

[7] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii, "Scalable K-means++," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 5, no. 7, pp. 622–633, 2012.

[8] Stephen V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77–89, 10 1997.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

ON THE GEOMETRIC-ARITHMETIC INDEX BY DECOMPOSITIONS

Juan C. Hernández[1], José M. Rodríguez[2] and José M. Sigarreta[1]

[1] *Facultad de Matemáticas, Universidad Autónoma de Guerrero*

[2] *Departmento de Matemáticas, Universidad Carlos III de Madrid*

emails: jcarloshg@gmail.com, jomaro@math.uc3m.es,
josemariasigarretaalmira@hotmail.com

Abstract

The concept of geometric-arithmetic index was introduced in the chemical graph theory recently, but it has shown to be useful. The main aim of this paper is to show that the computation of the geometric-arithmetic index of a graph G is essentially reduced to the computation of the geometric-arithmetic indices of the so-called primary subgraphs obtained by a decomposition of G. Furthermore, using these results, we can obtain formulas for the geometric-arithmetic indices of bridge graphs and other classes of graphs, like bouquet of graphs and circle graphs.

Key words: Graph invariant, Topological index, Geometric-Arithmetic index
MSC 2000: 05C07, 92E10

1 Introduction

A topological index is defined as a single number that represents a chemical structure in graph-theoretical terms via the molecular graph, and that correlates with a molecular property; it is used to understand physicochemical properties of chemical compounds. Topological indices are interesting since they capture some of the properties of a molecule in a single number. Hundreds of topological indices have been introduced and studied, starting with the seminal work by Wiener [23] in which he used the sum of all shortest-path distances of a (molecular) graph for modeling physical properties of alkanes.

Topological indices based on end-vertex degrees of edges have been used over 40 years. Among them, several indices are recognized to be useful tools in chemical researches. Probably, the best know such descriptor is the Randić connectivity index (R) [15]. There are

more than thousand papers and a couple of books dealing with this index (see, e.g., [7], [8], [16] and the references therein). During many years, scientists were trying to improve the predictive power of the Randić index. This led to the introduction of a large number of new topological descriptors resembling the original Randić index. The first geometric-arithmetic index GA_1, defined in [22] as

$$GA_1(G) = \sum_{uv \in E(G)} \frac{\sqrt{d_u d_v}}{\frac{1}{2}(d_u + d_v)},$$

where uv denotes the edge of the graph G connecting the vertices u and v, and d_u is the degree of the vertex u, is one of the successors of the Randić index. Although GA_1 was introduced just seven years ago, there are many papers dealing with this index (see, e.g., [5], [13], [17], [18], [22] and the survey [4]). There are other geometric-arithmetic indices, like $Z_{p,q}$ ($Z_{0,1} = GA_1$), but the results in [4, p.598] show empirically that the GA_1 index gathers the same information on observed molecules as other $Z_{p,q}$ indices.

The reason for introducing a new index is to gain prediction of some property of molecules somewhat better than obtained by already presented indices. Therefore, a test study of predictive power of a new index must be done. As a standard for testing new topological descriptors, the properties of octanes are commonly used. We can find 16 physico-chemical properties of octanes at www.moleculardescriptors.eu.

The graphic in [4, Fig.7] (from [4, Table 2], [21]) shows that there exists a good linear correlation between GA_1 and the heat of formation of benzenoid hydrocarbons (the correlation coefficient is equal to 0.972). Furthermore, the improvement in prediction with GA_1 index comparing to Randić index in the case of standard enthalpy of vaporization is more than 9%. Hence, one can think that GA_1 index should be considered in the QSPR/QSAR researches.

We say that a family of subgraphs $\{G_1, \ldots, G_r\}$ of G is a *decomposition* of G if $G_1 \cup \cdots \cup G_r = G$ and $G_i \cap G_j$ is either empty or a vertex for every $i, j \in \{1, \ldots, r\}, i \neq j$. The subgraphs $G_1, \ldots, G_r$ are usually called *primary subgraphs* of the decomposition.

Particular cases of decompositions are the T-decompositions (equivalent to the concept of graphs obtained by point-attaching, see [6], [19]); they are a very useful tool in different areas of graph theory, as Gromov hyperbolic graphs (see, e.g., [2], [3], [12]), polynomials on graphs (see [6]) and metric dimension of graphs (see [19]). We say that a vertex v of a graph G is a *cut-vertex* if $G \setminus \{v\}$ is not connected. Given a graph G, we say that a family of subgraphs $\{G_1, \ldots, G_r\}$ of G is a *T-decomposition* of G if $G_1 \cup \cdots \cup G_r = G$ and $G_i \cap G_j$ is either empty or a cut-vertex for every $i, j \in \{1, \ldots, r\}, i \neq j$. A graph is *biconnected* if it does not contain cut-vertices. The set of maximal biconnected components of a graph is an example of T-decomposition (the well-known biconnected decomposition of the graph); hence, every graph has a T-decomposition.

In [9] and [10] the authors introduce the concept of bridge and chain graphs, which

 ISBN: 978-84-608-6082-2

are particular cases of T-decompositions (a chain graph is a graph with a T-decomposition in which every cut-vertex belongs at most to two primary subgraphs; bridge graphs are a subset of chain graphs). For bridge and chain graphs the PI index was determined in [9] and for bridge graphs the Szeged index (and the vertex PI index) was considered in [11]. Recently, the Wiener index was considered in [1] for a class of graphs containing as special cases the bridge and chain graphs. Also, in [10] appear formulas for the Wiener, hyper-Wiener, detour and hyper-detour indices of bridge and chain graphs. Besides, [20] contains formulas for the Szeged, edge-Szeged, PI, vertex-PI and eccentric connectivity indices of splice graphs (another class of graphs with T-decompositions). More classes of graphs can be obtained as particular cases of T-decompositions: bouquets of graphs, rooted product graphs, corona product graphs and block graphs (see [19]).

Throughout this paper, $G = (V(G), E(G))$ denotes a (non-oriented) finite simple (without multiple edges and loops) connected graph with $E(G) \neq \emptyset$. Note that the connectivity of G is not an important restriction, since if G has connected components $G_1, \ldots, G_r$, then $GA_1(G) = GA_1(G_1) + \cdots + GA_1(G_r)$; furthermore, every molecular graph is connected.

2 Geometric-Arithmetic Index and decompositions

If $v \in V(G)$, then we denote by $N_G(v)$ or $N(v)$ the set of neighbors of v, i.e.,

$$N_G(v) = N(v) = \{u \in V(G) \mid uv \in E(G)\}.$$

Given a decomposition $\{G_1, \ldots, G_r\}$ of G, denote by $\mathcal{W}$ the set of vertices v in G belonging at least at two G_i's. Given a vertex $v \in \mathcal{W}$, denote by $G_{i_1}, \ldots, G_{i_k}$ the set of primary subgraphs containing v and by d_{i_j} the number of neighbors of v in G_{i_j} (then $d_v = d_{i_1} + \ldots + d_{i_k}$). If $v \in \mathcal{W}$, then we define $W(v)$ as

$$W(v) = \sum_{u \in N_G(v) \setminus \mathcal{W}} \frac{\sqrt{d_u d_v}}{\frac{1}{2}(d_u + d_v)} - \sum_{j=1}^{k} \sum_{u \in N_{G_{i_j}}(v) \setminus \mathcal{W}} \frac{\sqrt{d_u d_{i_j}}}{\frac{1}{2}(d_u + d_{i_j})}.$$

Denote by $\mathcal{Z}$ the set of edges in G with both endpoints in $\mathcal{W}$. If $e = uv \in \mathcal{Z}$, then $e \in G_i$ for some i, and we denote by d_u^*, d_v^* the degrees of u, v in G_i. If $e = uv \in \mathcal{Z}$, then we define $Z(e)$ as

$$Z(e) = \frac{\sqrt{d_u d_v}}{\frac{1}{2}(d_u + d_v)} - \frac{\sqrt{d_u^* d_v^*}}{\frac{1}{2}(d_u^* + d_v^*)}.$$

The following result allows to compute the precise value of $GA_1(G)$ in terms of the geometric-arithmetic indices of the primary subgraphs in any decomposition.

Theorem 2.1 *Let $\{G_1, \ldots, G_r\}$ be a decomposition of the graph G. Then*

$$GA_1(G) = \sum_{i=1}^{r} GA_1(G_i) + \sum_{v \in \mathcal{W}} W(v) + \sum_{e \in \mathcal{Z}} Z(e).$$

In order to estimate the difference between $GA_1(G)$ and $\sum_{i=1}^{r} GA_1(G_i)$, Proposition 2.2 below will provide bounds for $W(v)$ and $Z(uv)$.

Given a decomposition $\{G_1, \ldots, G_r\}$ of G and $e = uv \in \mathcal{Z}$, we say that e is *maximal* or *minimal* if $d_u = d_v$ or $d_u^* = d_v^*$, respectively.

Given a graph G, denote by Δ, δ the maximum and minimum degrees of G, respectively.

Proposition 2.2 *Let $\{G_1, \ldots, G_r\}$ be a decomposition of the graph G. Given $e \in \mathcal{Z}$, denote by Δ_e, δ_e the maximum and minimum degrees of the primary subgraph G_i with $e \in G_i$, respectively. Then*

$$-1 \leq \frac{2\sqrt{\Delta\delta}}{\Delta + \delta} - 1 \leq Z(e) \leq 1 - \frac{2\sqrt{\Delta_e \delta_e}}{\Delta_e + \delta_e} \leq 1,$$

for every $e \in \mathcal{Z}$. If e is maximal or minimal, then $Z(e) \geq 0$ or $Z(e) \leq 0$, respectively. Furthermore,

$$|W(v)| \leq \frac{1}{8}\sqrt{\Delta}\; d_v(d_v - 1),$$

for every $v \in \mathcal{W}$.

We have the following direct consequence of Theorem 2.1, since $|Z(e)| \leq 1$ for every $e \in \mathcal{Z}$ by Proposition 2.2.

Corollary 2.3 *Let $\{G_1, \ldots, G_r\}$ be a decomposition of the graph G. Then*

$$\Big| GA_1(G) - \sum_{i=1}^{r} GA_1(G_i) \Big| \leq \frac{1}{8}\,\Delta^{3/2}(\Delta - 1)\, card\, \mathcal{W} + card\, \mathcal{Z}.$$

Proposition 2.4 *Let $\{G_1, \ldots, G_r\}$ be a decomposition of the graph G. If $d_v \leq d_u$ for every $v \in \mathcal{W}$ and $u \in N_G(v) \setminus \mathcal{W}$, then*

$$GA_1(G) \geq \sum_{i=1}^{r} GA_1(G_i) - card\, \mathcal{Z}.$$

Furthermore, if every edge in $\mathcal{Z}$ is maximal, then

$$GA_1(G) \geq \sum_{i=1}^{r} GA_1(G_i).$$

Corollary 2.5 *Let $\{G_1, \dots, G_r\}$ be a decomposition of the graph G with minimum degree δ. If $d_v = \delta$ for every $v \in \mathcal{W}$, then*

$$GA_1(G) \geq \sum_{i=1}^{r} GA_1(G_i) - card\, \mathcal{Z}.$$

Furthermore, if every edge in $\mathcal{Z}$ is maximal, then

$$GA_1(G) \geq \sum_{i=1}^{r} GA_1(G_i).$$

Next, we apply our results on decompositions in order to compute the geometric-arithmetic indices of several chemical graphs.

Let $\{G_i\}_{i=1}^{d}$ be a set of finite pairwise disjoint graphs and $v_i, w_i \in V(G_i)$. The *chain graph*

$$C(G_1, G_2, \dots, G_d) = C(G_1, G_2, \dots, G_d; v_1, w_1, v_2, w_2, \dots, v_d, w_d)$$

of $\{G_i\}_{i=1}^{d}$ with respect to the vertices $\{v_i, w_i\}_{i=1}^{d}$ is the graph obtained from the graphs $G_1, \dots, G_d$ by identifying the vertex w_i and the vertex v_{i+1} for every $i \in \{1, 2, \dots, d-1\}$.

We denote by u_i the vertex in $C(G_1, G_2, \dots, G_d; v_1, w_1, \dots, v_d, w_d)$ obtained by identifying the vertices w_i and v_{i+1}. It is clear that $\{G_1, \dots, G_d\}$ is a T-decomposition of the chain graph:

$$C(G_1, \dots, G_d; v_1, w_1, \dots, v_d, w_d).$$

Following the notation in [9], given any graph H, $v, w \in V(H)$ and an integer $d > 1$, let us define

$$T_d(H, v, w) = C(H, \dots, H; v, w, \dots, v, w),$$

where H appears d times. Theorem 2.1 has the following direct consequence.

Corollary 2.6 *Consider any graph H, $v, w \in V(H)$ and an integer $d > 1$.*

(1) *If $vw \notin E(H)$, then*

$$GA_1(T_d(H, v, w)) = d\, GA_1(H) + (d-1)W(u_1).$$

(2) *If $vw \in E(H)$, then*

$$GA_1(T_d(H, v, w)) = d\, GA_1(H) + (d-1)W(u_1) + (d-2)Z(vw).$$

As an example, we deal with the Spiro chain of hexagons. Consider the chain graph $T_d(C_6)$, where C_6 is the cycle graph with vertices $v_1, v_2, \dots, v_6$ (labeled clockwise), $v = v_1$ and $w = v_4$. Since $vw \notin E(C_6)$, Corollary 2.6 gives

$$GA_1(T_d(C_6)) = d\, GA_1(C_6) + (d-1)W(u_1) = 6d + (d-1)\Big(4\,\frac{2\sqrt{8}}{6} - 4\Big) = \Big(2 + \frac{8\sqrt{2}}{3}\Big)d + 4 - \frac{8\sqrt{2}}{3}\,.$$

 ISBN: 978-84-608-6082-2

Note that if we take $v = v_3$ or $w = v_5$, then we obtain graphs with the same geometric-arithmetic index than the previous one.

Given any graph H, $v, w \in V(H)$, P_2 a graph of one edge connecting the two vertices v', w', and an integer $d > 1$, let us define

$$U_d(H) = C(H, P_2, \ldots, H, P_2, H; v, w, v', w', \ldots, v, w, v', w', v, w),$$

where H appears d times. Theorem 2.1 also has the following consequence.

Corollary 2.7 *Consider any graph H, $v, w \in V(H)$ and an integer $d > 1$. Denote by d_v, d_w the degrees of v, w in H, respectively.*

(1) *If $vw \notin E(H)$, then*

$$GA_1(U_d(H)) = d\,GA_1(H) + (d-1)\Big(W(u_1) + W(u_2) + \frac{2\sqrt{(d_v+1)(d_w+1)}}{d_v + d_w + 2}\Big).$$

(2) *If $vw \in E(H)$, then*

$$GA_1(U_d(H)) = d\,GA_1(H) + (d-1)\Big(W(u_1) + W(u_2) + \frac{2\sqrt{(d_v+1)(d_w+1)}}{d_v + d_w + 2}\Big)$$

$$+(d-2)Z(vw).$$

As another example, we deal with the Polyphenylenes. Consider the chain graph $U_d(C_6)$, where C_6 is the cycle graph with vertices $v_1, v_2, \ldots, v_6$ (labeled clockwise), $v = v_1$ and $w = v_4$. Since $vw \notin E(C_6)$, Corollary 2.7 gives

$$GA_1(U_d(C_6)) = d\,GA_1(C_6) + (d-1)\Big(W(u_1) + W(u_2) + \frac{2\sqrt{(d_v+1)(d_w+1)}}{d_v + d_w + 2}\Big)$$

$$= 6d + (d-1)\Big(2\Big(2\frac{2\sqrt{6}}{5} - 2\Big) + 1\Big) = \Big(3 + \frac{8\sqrt{6}}{5}\Big)d + 3 - \frac{8\sqrt{6}}{5}.$$

If we take $v = v_3$ or $w = v_5$, then we obtain graphs with the same geometric-arithmetic index than the previous one.

Let $\{G_i\}_{i=1}^d$ be a set of finite pairwise disjoint graphs and $v_i \in V(G_i)$. The *bridge graph*

$$B(G_1, G_2, \ldots, G_d) = B(G_1, G_2, \ldots, G_d; v_1, v_2, \ldots, v_d)$$

of $\{G_i\}_{i=1}^d$ with respect to the vertices $\{v_i\}_{i=1}^d$ is the graph obtained from the graphs $G_1, \ldots, G_d$ by connecting the vertices v_i and v_{i+1} by an edge for every $i \in \{1, 2, \ldots, d-1\}$. Theorem 2.1 has the following consequence.

Theorem 2.8 *Let $\{G_i\}_{i=1}^d$ be a set of finite pairwise disjoint graphs and $v_i \in V(G_i)$. Then*

$$GA_1(B(G_1,\dots,G_d;v_1,\dots,v_d)) = \sum_{i=1}^{d} GA_1(G_i) + d - 1 + \sum_{i=1}^{d} W(v_i) + \sum_{i=1}^{d-1} Z(v_iv_{i+1}).$$

Given any graph H, $v \in V(H)$ and an integer $d > 1$, let us define

$$G_d(H,v) = B(H,\dots,H;v,\dots,v),$$

where H appears d times. Theorem 2.8 has the following corollary.

Corollary 2.9 *Consider a graph H, $v \in V(H)$ and an integer $d > 1$. Denote by v_i the copy of v in the i-th copy of H in $G_d(H,v)$ and by d_v the degree of v in H.*

(1) *If $d = 2$, then*

$$GA_1(G_d(H,v)) = d\,GA_1(H) + 1 + 2W(v_1).$$

(2) *If $d \geq 3$, then*

$$GA_1(G_d(H,v)) = d\,GA_1(H) + d - 3 + 2W(v_1) + (d-2)W(v_2) +$$

$$+\frac{4\sqrt{(d_v+1)(d_v+2)}}{2d_v+3}.$$

As an example of the previous result, we consider $k \geq 3$, $d \geq 3$ and the bridge graph $G_d(P_k, w_1)$, where P_k is the path graph with vertices $w_1, w_2, \dots, w_k$. Corollary 2.9 gives

$$GA_1(G_d(P_k,w_1)) = d\,GA_1(P_k) + d - 3 + 2W(v_1) + (d-2)W(v_2) + \frac{4\sqrt{(d_v+1)(d_v+2)}}{2d_v+3}$$

$$= \Big(k - 3 + \frac{4\sqrt{2}}{3}\Big)d + d - 3 + 2\Big(1 - \frac{2\sqrt{2}}{3}\Big) + (d-2)\Big(\frac{2\sqrt{3}}{4} - \frac{2\sqrt{2}}{3}\Big) + \frac{4\sqrt{6}}{5}$$

$$= \Big(k - 2 + \frac{2\sqrt{2}}{3} + \frac{\sqrt{3}}{2}\Big)d - 1 - \sqrt{3} + \frac{4\sqrt{6}}{5}.$$

As another example, we deal with the Polyethene (when $d = 4$). For $d \geq 3$, consider the bridge graph $G_d(P_3, w_2)$. Corollary 2.9 gives

$$GA_1(G_d(P_3,w_2)) = d\,GA_1(P_3) + d - 3 + 2W(v_1) + (d-2)W(v_2) + \frac{4\sqrt{(d_v+1)(d_v+2)}}{2d_v+3}$$

 ISBN: 978-84-608-6082-2

$$= \frac{4\sqrt{2}}{3}\,d + d - 3 + 2\Big(2\Big(\frac{2\sqrt{3}}{4} - \frac{2\sqrt{2}}{3}\Big)\Big) + (d-2)\Big(2\Big(\frac{2\sqrt{4}}{5} - \frac{2\sqrt{2}}{3}\Big)\Big) + \frac{4\sqrt{12}}{7}$$

$$= \frac{13}{5}\,d - \frac{31}{5} + \frac{22\sqrt{3}}{7}\,.$$

The obtention of bounds for the geometric arithmetic index is a very active topic of research (see, e.g., [5], [13], [17], [22], the survey [4] and the references therein). In this section we are going to obtain an upper bound for it involving the numbers of vertices and the diameters of the primary subgraphs in the biconnected decomposition.

Let us define $M(n,r)$ as follows:

$$M(n,r) := r + \frac{1}{2}(n-r-1)(n-r+4).$$

We denote by $diam\,(G)$ the diameter of G: $diam\,(G) := \max\{d(u,v) \,|\, u,v \in V(G)\}$.

Theorem 2.10 *If $\{G_1,\ldots,G_k\}$ is the biconnected decomposition of a graph G and G_j has n_j vertices for $1 \le j \le k$, then*

$$GA_1(G) \le \sum_{j=1}^{k} M(n_j, diam\,(G_j)).$$

Acknowledgements

Supported in part by a grant from CONACYT (FOMIX-CONACyT-UAGro 249818), México, and two grants from Ministerio de Economía y Competitividad (MTM 2013-46374-P and MTM 2015-69323-REDT), Spain.

References

[1] R. Balakrishnan, N. Sridharan and K. Viswanathan Iyer, *Wiener index of graphs with more than one cut-vertex*, Appl. Math. Lett. **21** (2008) 922–927.

[2] S. Bermudo, J. M. Rodríguez, J. M. Sigarreta and J.-M. Vilaire, *Gromov hyperbolic graphs*, Discr. Math. **313** (2013) 1575–1585.

[3] G. Brinkmann, J. Koolen and V. Moulton, *On the hyperbolicity of chordal graphs*, Ann. Comb. **5** (2001) 61–69.

[4] K. C. Das, I. Gutman and B. Furtula, *Survey on Geometric-Arithmetic Indices of Graphs*, MATCH Commun. Math. Comput. Chem. **65** (2011) 595–644.

[5] K. C. Das, I. Gutman and B. Furtula, *On first geometric-arithmetic index of graphs*, Discrete Appl. Math. **159** (2011) 2030–2037.

[6] E. Deutsch and S. Klavžar, *Computing Hosoya polynomials of graphs from primary subgraphs*, MATCH Commun. Math. Comput. Chem. **70** (2013) 627–644.

[7] I. Gutman and B. Furtula, *Recent Results in the Theory of Randić Index*, Univ. Kragujevac, Kragujevac, 2008.

[8] X. Li and I. Gutman, *Mathematical Aspects of Randić Type Molecular Structure Descriptors*, Univ. Kragujevac, Kragujevac, 2006.

[9] T. Mansour and M. Schork, *The PI index of bridge and chain graphs*, MATCH Commun. Math. Comput. Chem. **61** (2009) 723–734.

[10] T. Mansour and M. Schork, *Wiener, hyper-Wiener, detour and hyper-detour indices of bridge and chain graphs*, J. Math. Chem. **581** (2009) 59–69.

[11] T. Mansour and M. Schork, *The vertex PI index and Szeged index of bridge graphs*, Discrete Appl. Math. **157** (2009) 1600–1606.

[12] J. Michel, J. M. Rodríguez, J. M. Sigarreta and V. Villeta, *Hyperbolicity and parameters of graphs*, Ars Comb. **100** (2011) 43–63.

[13] M. Mogharrab and G. H. Fath-Tabar, *Some bounds on GA_1 index of graphs*, MATCH Commun. Math. Comput. Chem. **65** (2010) 33–38.

[14] O. Ore, *Diameters in Graphs*, J. Comb. Theory **5** (1968) 75–81.

[15] M. Randić, *On characterization of molecular branching*, J. Am. Chem. Soc. **97** (1975) 6609–6615.

[16] J. A. Rodríguez and J. M. Sigarreta, *On the Randić index and condicional parameters of a graph*, MATCH Commun. Math. Comput. Chem. **54** (2005) 403–416.

[17] J. M. Rodríguez and J. M. Sigarreta, *On the Geometric-Arithmetic Index*, MATCH Commun. Math. Comput. Chem. **74** (2015) 103–120.

[18] J. M. Rodríguez and J. M. Sigarreta, *Spectral properties of the Geometric-Arithmetic Index*, Appl. Math. Comput. **277** (2016) 142–153.

[19] J. A. Rodríguez-Velázquez, C. García Gómez and G. A. Barragán-Ramírez, *Computing the local metric dimension of a graph from the local metric dimension of primary subgraphs*, Int. J. Comput. Math. **92(4)** (2015) 686–693.

 ISBN: 978-84-608-6082-2

[20] R. Sharafdini and I. Gutman, *Splice graphs and their topological indices*, Kragujevac J. Sci. **35** (2013) 89–98.

[21] *TRC Thermodynamic Tables. Hydrocarbons*, Thermodynamic Research Center, The Texas A & M University System: College Station, TX, 1987.

[22] D. Vukičević and B. Furtula, *Topological index based on the ratios of geometrical and arithmetical means of end-vertex degrees of edges*, J. Math. Chem. **46** (2009) 1369–1376.

[23] H. Wiener, *Structural determination of paraffin boiling points*, J. Am. Chem. Soc. **69** (1947) 17–20.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Solving algebraic Riccati equations with an efficient iterative process with fourth order of convergence

M.A. Hernández-Verón [1] **and N. Romero**[1]

[1] *Department of Mathematics and Computation, University of La Rioja*

emails: mahernan@unirioja.es, natalia.romero@unirioja.es

Abstract

Key words: Iterative methods; Algebraic Riccati equations; Semilocal and Local convergence.

1 Introduction

The prototype of an algebraic Riccati equation [9] is given by the quadratic matrix equation of the form

$$\mathcal{R}(X) := XDX - XA - BX - C = 0,$$

where the coefficients A, B, C, D are real or complex $n \times n$ matrices and $n \times n$ matrix solutions X are to be found. The growth of interest in algebraic Riccati equations in the last years has been explosive. Primarily, this has been driven by the important role played by these equations in optimal filter design and control theory. Moreover, they may arise in many areas of scientific computing and engineering applications such as the total least squares, problems with or without symmetric constraints, the transport theory, the Wiener-Hopf factorization of Markov chains, the optimal solutions of linear differential systems, for instance see [8, 5].

Frequently, algebraic Riccati equation take a symmetric form:

$$\mathcal{R}(X) := XDX - XA - A^*X - C = 0, \tag{1}$$

where C and D are hermitian matrices (A^* is the complex conjugate of the transpose of A).

The main concern of this work is with the numerical hermitian solutions of these algebraic Riccati equation (1), which are required for physical reasons. Firstly, we consider the

symmetric algebraic Riccati equation (1) under the hypotheses that $D = D^*$, $C = C^*$, the pair (A, D) is stabilizable (if there is a feedback matrix K such that $A + BK$ is stable, that is its eigenvalues are all in the open left half-plane), and the size of matrix A, D and C is $n \times n$. Under these conditions, the existence of hermitian solutions X of $\mathcal{R}(X) = 0$ can be characterized using spectral properties of the matrix $\begin{pmatrix} -A & D \\ C & A^* \end{pmatrix}$, see [9].

The Riccati function $\mathcal{R}(X) = XDX - XA - A^*X - C$ that we consider maps hermitian matrices to hermitian matrices. The set of all hermitian matrices of size $n \times n$ forms a linear vector space $\mathcal{H}$ over $\mathbb{R}$, and it is possible to formulate the Frechét derivatives of the function $\mathcal{R}$. The first Frechét derivatives at a matrix X is a linear map $\mathcal{R}'(X) : \mathcal{H} \to \mathcal{H}$ and is easily found to be ([13]):

$$\mathcal{R}'(X)E = E(DX - A) + (DX - A)^*E. \tag{2}$$

Also the second derivative at X, $\mathcal{R}'(X) : (\mathcal{H} \times \mathcal{H}) \to \mathcal{H}$ is given by

$$\mathcal{R}''(X)E_1E_2 = E_1DE_2 + E_2DE_1 \tag{3}$$

therefore, the second derivative is a bilinear constant operator.

Although algebraic Riccati equations (1) are truly nonlinear equation, many studies have been made to solve such equations, by methods which rely heavily on linear algebra and the theory of matrices, see [9, 14]. However, our main objective is approximate a solution of the algebraic Riccati equation (1) using iterative methods ([13]). Basically the iterative process used is the well known Newton-Kantorovich method [12]:

$$\begin{cases} X_0 \text{ given,} \\ X_{n+1} = X_n - [\mathcal{R}'(X_n)]^{-1}\mathcal{R}(X_n), \quad n \geq 0, \end{cases} \tag{4}$$

or variants of this method, see for example [3, 7]. Notice that, apply the Newton-Kantorovich method to approximate a solution of algebraic Riccati equation (1) is simply to solve the Lyapunov equation:

$$X_{n+1}(A - DX_n) + (A - DX_n)^*X_{n+1} = -H(X_n) - C,$$

where H is the operator $H(X) = XDX$. The Lyapunov equation

$$SA + A^*S = W, \qquad \text{where } A, W \text{ are given and } W \text{ is hermitian,} \tag{5}$$

is the most common problem in the class of problems which involve matrix equations. If Lyapunov equation is solved as a set of $n(n+1)/2$ equations in $n(n+1)/2$ variables the opertional cost is $\mathcal{O}(n^6)$ operations. However, there exist fast methods, that exploit the special structure of the linear equations, based on first reduce A to Schur or upper

Hessenberg form. For instance, the Bartels-Stewart algorithm [4] for the Lyapunov equation has operational cost $\mathcal{O}(n^3)$ operations. The Bartels-Stewart algorithm is one of the most effective schemes for solving the symmetric linear matrix equation. The algorithm is ideally suited to the sequential solution of (5) with different W, and with the same A matrix.

Different results were obtained approximating the solution of (1) by iterative processes with quadratic convergence. In this paper, we consider the possibility of doing so through an iterative process with higher order of convergence. The iterative process with cubical convergence most commonly used is the Chebyshev method [6]:

$$\begin{cases} X_0 \text{ given}, \\ Y_n = X_n - [\mathcal{R}'(X_n)]^{-1}\mathcal{R}(X_n), \\ X_{n+1} = Y_n + \frac{1}{2}L_{\mathcal{R}}(X_n)(Y_n - X_n), \quad n \geq 0, \end{cases}$$

where $L_{\mathcal{R}}(X_n) = [\mathcal{R}'(X_n)]^{-1}\mathcal{R}''(X_n)[\mathcal{R}'(X_n)]^{-1}\mathcal{R}(X_n)$. In this case note that, apply the Chebyshev method to approximate a solution of algebraic Riccati equation (1) is reduced to solve two Lyapunov equations:

$$Y_n(A - DX_n) + (A - DX_n)^* Y_n = -H(X_n) - C,$$

$$X_{n+1}(A - DX_n) + (A - DX_n)^* X_{n+1} = H(Y_n - X_n) - H(X_n) - C,$$

Notice that, the difference with the application of the Newton-Kantorovich method is one more Lyapunov equation to solve and to obtain the new matrix $H(Y_n - X_n)$. Of course, higher order schemes require more computational cost than other simpler methods, which makes them unfavorable to be used in general. However, the existence of an extensive literature of high order methods [1, 2] reveals they are only limited by the nature of the problem to be solved. In this paper, we construct a parameter family of iterative processes with at least fourth order of convergence to approximate a solution of (1). Thus, for a fixed value of the parameter, it is obtained the fourth efficient iterative process:

$$\begin{cases} X_0 \quad \text{given}, \\ Y_n = X_n - [\mathcal{R}'(X_n)]^{-1}\mathcal{R}(X_n), \\ Z_n = Y_n + \dfrac{1}{2}[\mathcal{R}'(X_n)]^{-1}\left[\dfrac{3}{2}(\mathcal{R}'(X_n) - \mathcal{R}'(X_n + \dfrac{2}{3}(Y_n - X_n)))\right](Y_n - X_n), \quad n \geqslant 0, \\ X_{n+1} = Z_n + \dfrac{1}{2}\left[\dfrac{3}{2}[\mathcal{R}'(X_n)]^{-1}(\mathcal{R}'(X_n) - \mathcal{R}'(X_n + \dfrac{2}{3}(Y_n - X_n)))\right]^2(Y_n - X_n), \quad n \geqslant 0. \end{cases} \tag{6}$$

Iterative process (6) only involve the Riccati function and its first Frechét derivative as the Newton method. And therefore, the operational cost is similar to the Newton method. Thus, it is an efficient iterative method.

We show that apply iterative process (6) to approximate a solution of algebraic Riccati equation (1) is simply to solve three Lyapunov equations. In fact, the three Lyapunov equations are:

$$\begin{aligned} Y_n(A-DX_n)+(A-DX_n)^*Y_n &= -H(X_n)-C, \\ Z_n(A-DX_n)+(A-DX_n)^*Z_n &= -H(X_n)-C+H(Y_n-X_n). \\ X_{n+1}(A-DX_n)+(A-DX_n)^*X_{n+1} &= -H(X_n)-C+(Y_n-X_n)D(Z_n-Y_n) \\ &\quad +(Z_n-Y_n)D(Y_n-X_n). \end{aligned}$$

Therefore, the difference with the application of the Newton-Kantorovich method is two more Lyapunov equation to solve and to obtain the new matrices $(Y_n - X_n)D(Z_n - Y_n)$, $(Z_n - Y_n)D(Y_n - X_n)$, $H(Y_n - X_n)$. Note that, matrix $H(Y_n - X_n)$, $(Y_n - X_n)D$ and $D(Y_n - X_n)$ are already computed when we apply the Chebyshev method. Thus, this method has good numerical behavior.

Acknowledgements

The research has been partially supported by the project MTM2014-52016-C2-1-P of the Spanish Ministry of Economy and Competitiveness.

References

[1] G. Altman, Iterative methods of higher order, Bull. Acad. Pollon. Sci. (Série des Sci., Math., Astr., Phys. IX (1961) 62–68.

[2] S. Amat, M. A. Hernández and N. Romero, A modified Chebyshev's iterative method with at least six-th order of convergence, Appl. Math. Comput., 206 (2008) 164– 174.

[3] Z.Z. Bai, X.X. Guo and J.F. Yin, *On two iteration methods for the quadratic matrix equations*, Int. J. Numer. Anal. Mod., 2, (2005), 114–122.

[4] R.H. Bartels and G.W. Stewart, Algorithm 432: Solution of the matrix equation $AX + XB = C$ Commun. Ass. Comput. Mach., 15 (9) (1972), 820–826.

[5] J.E. Dennis and R.B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations; *SIAM*; Philadelphia, 1996.

[6] M. A. Hernández, Chebyshev's approximation algorithms and applications, Comput. Math. Appl., 41 (2001) 433–445.

[7] N.J. Higham and H.M. Kim, *Solving a quadratic matrix equation By Newton's method with exact line searches*, SIAM J. Matrix Anal. Appl., 23, (2001), 303–316.

[8] P. Lancaster and M. Tismenetsky, The theory of matrices with applications; *Accademic Press*; Orlando, 1985.

[9] P. Lancaster and L. Rodman, Algebraic Riccati equations; *Oxford Science Publications*; Oxford, 1995.

[10] L.C.G. Rogers, *Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains*, Ann. Appl. Probab., 4, (1994), 390–413.

[11] L.C.G. Rogers and Z. Shi, *Computing the invariant law of a fluid model*, J. Appl. Probab., 31, (1994), 885–896.

[12] L. V. Kantorovich and G. P. Akilov, *Functional analysis*, Pergamon Press, Oxford, 1982.

[13] J.M. Ortega and W.C. Rheinboldt *Iterative solution of nonlinear equations in several variables*, Accademic Press, New York, 1970.

[14] R.S. Varga, *Matrix Iterative Analysis*, 2nd ed. Springer, Berlin, 2000.

 ISBN: 978-84-608-6082-2

Volume III

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Serial concatenation of a block code and a 2D convolutional code

V. Herranz[1] and C. Perea[1]

[1] *Center of Operations Research. Statistics, Mathematics and Informatics Department. Miguel Hernndez University, University of First and Second Authors*

emails: mavi.herranz@umh.es, perea@umh.es

Abstract

In this work we introduce an input-state-output representation of four models of concatenation of a block code and a 2D convolutional code.

Key words: 2D convolutional code, block code, input-state-output representation, linear systems, concatenated code.

1 Introduction

The concatenation of an outer block code and an inner convolutional code were introduced by Forney [3] as a way of providing long codes with a manageable decoding complexity.

On the other hand, in recent years, multidimensional (mD) convolutional codes has gained interest as an alternative to 1D convolutional codes as working on a framework that take advantage of the correlation of data in several dimensions of time. These codes have many applications, starting from transmition of m-dimensional data, such a 2D pictures, 3D animatation, etc. (see [5, 8]), through to the storage of digital information [7]. While 1D convolutional codes have been thoroughtly studied, the literature about mD convolutional codes is quite limited. Gluesing-Luerssen, Rosenthal and Weiner analyze the relation between multidimensional convolutional codes and systems [4]. More recently, Napp, Perea and Pinto [6], introduce the natural extension of the input-state-output representation of 1D convolutional codes to 2D convolutional codes.

In this work we present the natural, although non-trivial, generalization of the input-state-output representation of concatenation of a block code and a 1D convolutional code

introduced in Climent, Herranz and Perea [1] to the input-state-output representation of a block code and a 2D convolutional code.

The paper is structured as follows. Our starting point in Section 2 will be linear systems representation of 2D convolutional codes, that we will need throughout the paper. In Section 3 we characterize four models of concatenation of a block code and a 2D convolutional from a linear systems theory viewpoint.

2 Preliminary results

Denote by $\mathbb{F} = GF(q)$ the Galois field of q elements and $\mathbb{F}[z_1, z_2]$ the ring of polynomials in 2 indeterminates with coefficients in $\mathbb{F}$.

Error correcting codes are usually divided into two distinct classes: block codes and convolutional codes. A block code is the primary type of channel coding which was used in earlier mobile communication systems.

A *block code* $\mathcal{B}$ can be characterized as the set of vectors $\boldsymbol{v} \in \mathbb{F}^n$ such that $\boldsymbol{v} = G\boldsymbol{u}$ for some $\boldsymbol{u} \in \mathbb{F}^k$ (with $k \in \mathbb{N}$), where G is an $n \times k$ matrix of full column rank that we call a *generator matrix* of $\mathcal{B}$. In addition, if $G = \begin{pmatrix} I_k \\ P \end{pmatrix}$, we say that G is *systematic*. Also, we say that $\boldsymbol{u}$ is the *information vector* and that $\boldsymbol{v}$ is the corresponding code vector or *codeword*. We refer $\mathcal{B}$ as an $[n, k]$-code.

Next we give some preliminaries on 2D linear systems, which we will use to construct 2D finite support convolutional codes. In particular, we consider the Fornasini-Marchesini state space model representation of 2D systems [2]. In this model a first quarter plane 2D linear system, denoted by

$$\Sigma = (A_1, A_2, B_1, B_2, C, D),$$

is given by the updating equations

$$\left.\begin{array}{rcccccccc} \boldsymbol{x}(i+1,j+1) & = & A_1\boldsymbol{x}(i,j+1) & + & A_2\boldsymbol{x}(i+1,j) & + & B_1\boldsymbol{u}(i,j+1) & + & B_2u(i+1,j) \\ \boldsymbol{y}(i,j) & = & C\boldsymbol{x}(i,j) & + & D\boldsymbol{u}(i,j) & & & & \end{array}\right\}, \tag{1}$$

where $A_1, A_2 \in \mathbb{F}^{\delta\times\delta}$, $B_1, B_2 \in \mathbb{F}^{\delta\times k}$, $C \in \mathbb{F}^{(n-k)\times\delta}$, $D \in \mathbb{F}^{(n-k)\times k}$, δ, n, $k \in \mathbb{N}$, $n > k$ and with past finite support of the input and of the state and zero initial conditions (i.e., $\boldsymbol{u}(i,j) = \boldsymbol{x}(i,j) = 0$ for $i < 0$ or $j < 0$ and zero initial conditions $\boldsymbol{x}(0,0) = 0$. We say that $\Sigma = (A_1, A_2, B_1, B_2, C, D)$ has dimension δ, local state $\boldsymbol{x}(i,j)$, input $\boldsymbol{u}(i,j)$ and output $\boldsymbol{y}(i,j)$ at (i,j).

We denote by $\mathcal{C}(A_1, A_2, B_1, B_2, C, D)$ the 2D finite suppport convolutional code whose codewords are the finite-weight input-output trajectories of the system $\Sigma = (A_1, A_2, B_1, B_2, C, D)$. Moreover, Σ is called an *input-state-output (ISO) representation* of $\mathcal{C}(A_1, A_2, B_1, B_2, C, D)$.

 ISBN: 978-84-608-6082-2

3 Models of concatenation of a block and a 2D convolutional code

In this section we introduce four models of concatenation of a block code and a 2D convolutional code. Let $\mathcal{C}_o$ be a block code, that we call *outer encoder*, and let $\mathcal{C}_i$ be a 2D convolutional code, that we call *inner encoder*.

Let G_o be an encoder of the block code $\mathcal{C}_o$ and let $\Sigma_2 = (A_1^{(2)}, A_2^{(2)}, B_1^{(2)}, B_2^{(2)}, C^{(2)}, D^{(2)})$ be an ISO representation of the 2D convolutional code $\mathcal{C}_i$.

Let $\boldsymbol{u}^1(i,j)$ be the information vector of $\mathcal{C}_o$, and let $\boldsymbol{x}^2(i,j)$, $\boldsymbol{u}^2(i,j)$, and $\boldsymbol{y}^2(i,j)$ be the state vector, the information vector and the parity vector of $\mathcal{C}_i$, respectively. We will consider the interconnection models defined in [1] for 1D systems.

3.1 First model

In the first model of concatenation, the outer encoder $\mathcal{C}_o$ and the inner encoder $\mathcal{C}_i$ are serialized, one after the other as described in [1] for 1D systems (see Figure 1). The input information $\boldsymbol{u}^1(i,j)$ is fed to $\mathcal{C}_o$ and the obtained codeword

$$\boldsymbol{v}^1(i,j) = G_o \boldsymbol{u}^1(i,j) \tag{2}$$

is then encoded by $\mathcal{C}_i$ in a way that

$$\boldsymbol{u}^2(i,j) = \boldsymbol{v}^1(i,j) \tag{3}$$

So, the codewords $\boldsymbol{v}^2(i,j)$ of $\mathcal{C}_i$ are given by

$$\boldsymbol{v}^2(i,j) = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ \boldsymbol{u}^2(i,j) \end{pmatrix} \tag{4}$$

We denote by $\mathcal{BC}^{(1)}$ the corresponding 2D concatenated convolutional code. Observe that the vector state $\boldsymbol{x}(i,j)$, the information vector $\boldsymbol{u}(i,j)$ and the parity vector $\boldsymbol{y}(i,j)$ of $\mathcal{BC}^{(1)}$ are given by

$$\boldsymbol{x}(i,j) = \boldsymbol{x}^2(i,j), \qquad \boldsymbol{u}(i,j) = \boldsymbol{u}^1(i,j), \qquad \text{and} \qquad \boldsymbol{y}(i,j) = \boldsymbol{y}^2(i,j). \tag{5}$$

So the codewords $\boldsymbol{v}(i,j)$ of $\mathcal{BC}^{(1)}$ are given by

$$\boldsymbol{v}(i,j) = \begin{pmatrix} \boldsymbol{y}(i,j) \\ \boldsymbol{u}(i,j) \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ \boldsymbol{u}^1(i,j) \end{pmatrix}. \tag{6}$$

Next theorem introduces an input-state-output representation of the 2D concatenated convolutional code $\mathcal{BC}^{(1)}$ from the generator matrix of the outer code $\mathcal{C}_o$ and an input-state-output representation of the inner code $\mathcal{C}_i$.

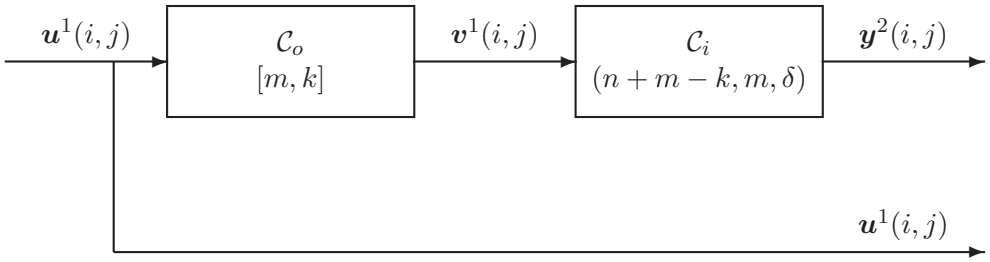


Figure 1: Concatenated code $\mathcal{BC}^{(1)}$

Theorem 1: *Let $\mathcal{C}_o$ be an $[m,k]$-code with generator matrix G_o and let $\mathcal{C}_i(A_1^{(2)}, A_2^{(2)}, B_1^{(2)}, B_2^{(2)}, C^{(2)}, D^{(2)})$ be an $(n+m-k,m,\delta)$-2D convolutional code . Then an input-state-output representation for the rate k/n 2D concatenated code $\mathcal{BC}^{(1)}$ is given by system (1), with*

$$A_1 = A_1^{(2)}, \quad A_2 = A_2^{(2)}, \quad B_1 = B_1^{(2)}G_o, \quad B_2 = B_2^{(2)}G_o, \quad C = C^{(2)}, \quad \textit{and} \quad D = D^{(2)}G_o. \tag{7}$$

3.2 Second model

If in the first model we consider that the generator matrix G_o of the block code $\mathcal{C}_o$ is in systematic form, that is, $G_o = \begin{pmatrix} I_k \\ P_o \end{pmatrix}$ and we consider the parity check vector $P_o\boldsymbol{u}^1(i,j)$ of the obtained codeword

$$\boldsymbol{v}^1(i,j) = \begin{pmatrix} \boldsymbol{u}^1(i,j) \\ P_o\boldsymbol{u}^1(i,j) \end{pmatrix} \tag{8}$$

of $\mathcal{C}_o$ as the input of $\mathcal{C}_i$, that is,

$$\boldsymbol{u}^2(i,j) = P_o\boldsymbol{u}^1(i,j), \tag{9}$$

then we obtain the 2D concatenated code of Figure 2. Here, the codewords $\boldsymbol{v}^2(i,j)$ of $\mathcal{C}_i$ are given by

$$\boldsymbol{v}^2(i,j) = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ \boldsymbol{u}^2(i,j) \end{pmatrix}. \tag{10}$$

We denote by $\mathcal{BC}_{\text{sys}}^{(1)}$ the corresponding 2D concatenated convolutional code. Analogously to the model $\mathcal{BC}^{(1)}$, the vector state $\boldsymbol{x}(i,j)$, the information vector $\boldsymbol{u}(i,j)$, the parity

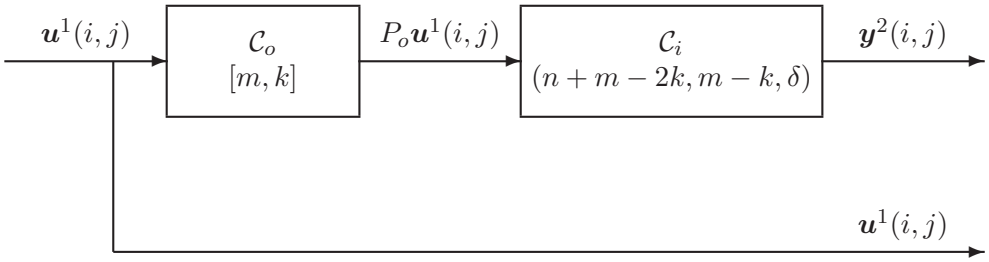


Figure 2: Concatenated code $\mathcal{BC}^{(1)}_{\rm sys}$

vector $\boldsymbol{y}(i,j)$ and the codewords $\boldsymbol{v}(i,j)$ of $\mathcal{BC}^{(1)}_{\rm sys}$ are given by expressions (5) and (6). An input-state-output representation of the 2D concatenated convolutional code $\mathcal{BC}^{(1)}_{\rm sys}$ is given by the following theorem.

Theorem 2: *Let $\mathcal{C}_o$ be an $[m,k]$-code with generator matrix $G_o = \begin{pmatrix} I_k \\ P_o \end{pmatrix}$ and let $\mathcal{C}_i(A_1^{(2)}, A_2^{(2)}, B_1^{(2)}, B_2^{(2)}, C^{(2)}, D^{(2)})$ be an $(n+m-2k, m-k, \delta)$-2D convolutional code. Then an input-state-output representation for the rate k/n 2D concatenated convolutional code $\mathcal{BC}^{(1)}_{sys}$ is given by system (1), with*

$$A_1 = A_1^{(2)}, \quad A_2 = A_2^{(2)}, \quad B_1 = B_1^{(2)}P_o, \quad B_2 = B_2^{(2)}P_o, \quad C = C^{(2)}, \quad \text{and} \quad D = D^{(2)}P_o. \tag{11}$$

3.3 Third model

If we vary the code $\mathcal{BC}^{(1)}$ so that the codeword $\boldsymbol{v}^1(i,j)$ of the outer encoder is a part of the concatenated codeword (see Figure 3), we obtain a new 2D concatenated convolutional code that we denote by $\mathcal{BC}^{(2)}$.

Here, the information vector $\boldsymbol{u}^1(i,j)$ and the codeword $\boldsymbol{v}^1(i,j)$ of $\mathcal{C}_o$ and the information vector $\boldsymbol{u}^2(i,j)$ and the codewords $\boldsymbol{v}^2(i,j)$ of $\mathcal{C}_i$ are the same as in the first model (see expressions (2), (3) and (4)). Then, the vector state $\boldsymbol{x}(i,j)$, the information vector $\boldsymbol{u}(i,j)$ and the parity vector $\boldsymbol{y}(i,j)$ of $\mathcal{BC}^{(2)}$ are given by

$$\boldsymbol{x}(i,j) = \boldsymbol{x}^2(i,j), \quad \boldsymbol{u}(i,j) = \boldsymbol{u}^1(i,j), \quad \text{and} \quad \boldsymbol{y}(i,j) = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ \boldsymbol{v}^1(i,j) \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ G_o\boldsymbol{u}^1(i,j) \end{pmatrix} \tag{12}$$

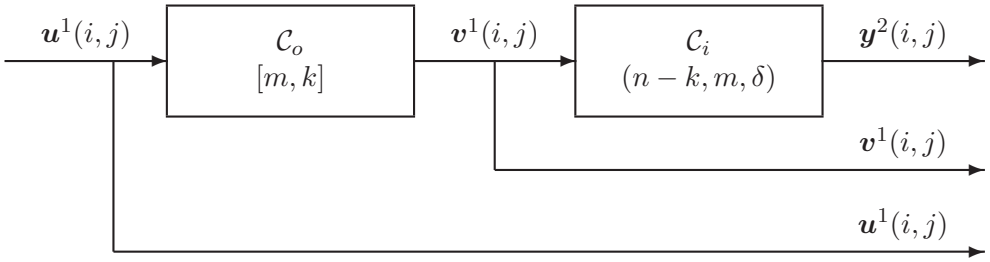


Figure 3: Concatenated code $\mathcal{BC}^{(2)}$

and the codewords $\boldsymbol{v}(i,j)$ of $\mathcal{BC}^{(2)}$ are given by

$$\boldsymbol{v}(i,j) = \begin{pmatrix} \boldsymbol{y}(i,j) \\ \boldsymbol{u}(i,j) \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ G_o \boldsymbol{u}^1(i,j) \\ \boldsymbol{u}^1(i,j) \end{pmatrix}. \tag{13}$$

Next theorem provides an input-state-output representation of the 2D convolutional code $\mathcal{BC}^{(2)}$ from the generator matrix of the outer encoder $\mathcal{C}_o$ and an input-state-output representation of the inner encoder $\mathcal{C}_i$.

Theorem 3: *Let $\mathcal{C}_o$ be an $[m,k]$-code with generator matrix G_o and let $\mathcal{C}_i(A_1^{(2)}, A_2^{(2)}, B_1^{(2)}, B_2^{(2)}, C^{(2)}, D^{(2)})$ be an $(n-k,m,\delta)$-2D convolutional code. Then an input-state-output representation for the rate k/n 2D concatenated convolutional code $\mathcal{BC}^{(2)}$ is given by system (1), with*

$$\begin{gathered} A_1 = A_1^{(2)}, \quad A_2 = A_2^{(2)}, \quad B_1 = B_1^{(2)} G_o, \quad B_2 = B_2^{(2)} G_o \\ C = \begin{pmatrix} C_2 \\ O \end{pmatrix}, \quad D = \begin{pmatrix} D_2 G_o \\ G_o \end{pmatrix}. \end{gathered} \tag{14}$$

3.4 Fourth model

Finally, if we make to $\mathcal{BC}^{(2)}$ a similar variation as the variation made to $\mathcal{BC}^{(1)}$ in order to get $\mathcal{BC}^{(1)}_{\text{sys}}$, we obtain a 2D concatenated convolutional code that we denote by $\mathcal{BC}^{(2)}_{\text{sys}}$ (see Figure 4).

Here, the information vectors $\boldsymbol{u}^1(i,j)$, $\boldsymbol{u}^2(i,j)$ and the codewords $\boldsymbol{v}^1(i,j)$, $\boldsymbol{v}^2(i,j)$ of $\mathcal{C}_o$ and $\mathcal{C}_i$, respectively, are the same as in $\mathcal{BC}^{(1)}_{\text{sys}}$ (see expressions (8), (9) and (10)). Also,

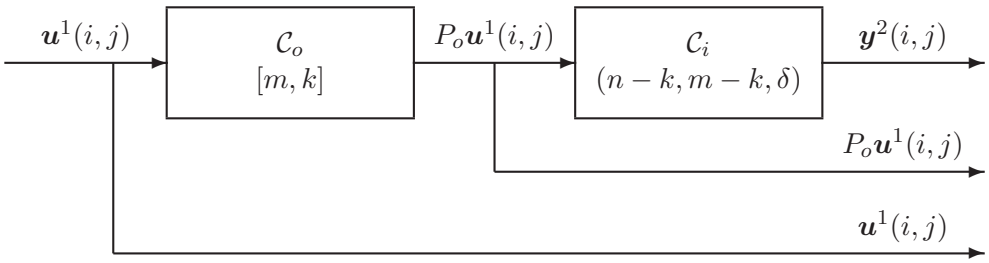


Figure 4: Concatenated code $\mathcal{BC}^{(2)}_{\rm sys}$

the state vector $\boldsymbol{x}(i,j)$ and the information vector $\boldsymbol{u}(i,j)$ of $\mathcal{BC}^{(2)}_{\rm sys}$ are the same as in $\mathcal{BC}^{(1)}_{\rm sys}$ (see expression (5) and the comments before Theorem 2). But in the 2D concatenated code $\mathcal{BC}^{(2)}_{\rm sys}$, the parity vector $\boldsymbol{y}(i,j)$ and the codeword $\boldsymbol{v}(i,j)$ are given by

$$\boldsymbol{y}(i,j) = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ P_o\boldsymbol{u}^1(i,j) \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{v}(i,j) = \begin{pmatrix} \boldsymbol{y}(i,j) \\ \boldsymbol{u}(i,j) \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}^2(i,j) \\ P_o\boldsymbol{u}^1(i,j) \\ \boldsymbol{u}^1(i,j) \end{pmatrix}. \tag{15}$$

So, an input-state-output representation of the 2D concatenated code $\mathcal{BC}^{(2)}_{\rm sys}$ is given by the following theorem.

Theorem 4: *Let $\mathcal{C}_o$ be an $[m,k]$-code with generator matrix in systematic form $G_o = \begin{pmatrix} I_k \\ P_o \end{pmatrix}$ and let $\mathcal{C}_i(A_1^{(2)}, A_2^{(2)}, B_1^{(2)}, B_2^{(2)}, C^{(2)}, D^{(2)})$ be an $(n-k, m-k, \delta)$-2D convolutional code. Then an input-state-output representation for the rate k/n 2D concatenated convolutional code $\mathcal{BC}^{(2)}_{sys}$ is given by system (1), with*

$$\begin{gathered} A_1 = A_1^{(2)}, \quad A_2 = A_2^{(2)}, \quad B_1 = B_1^{(2)}P_o, \quad B_2 = B_2^{(2)}P_o \\ C = \begin{pmatrix} C_2 \\ O \end{pmatrix}, \quad D = \begin{pmatrix} D_2P_o \\ P_o \end{pmatrix}. \end{gathered} \tag{16}$$

References

[1] Joan-Josep Climent, Victoria Herranz, and Carmen Perea. Linear system modelization of concatenated block and convolutional codes. *Linear Algebra and its Applications*, 429:1191–1212, 2008.

[2] Ettore Fornasini and Giovanni Marchesini. Structure and properties of two-dimensional systems. In S. G. Tzafestas, editor, *Multidimensional Systems, Techniques and Applications*, pages 37–88, 1986.

[3] G. David Forney, Jr. *Concatenated Codes*. MIT Press, Cambridge, MA, 1966.

[4] Heide Gluesing-Luerssen, Joachim Rosenthal, and Paul A. Weiner. Duality between multidimensional convolutional codes and systems. In F. Colonius, U. Helmke, F. Wirth, and D. Prätzel-Wolters, editors, *Advances in Mathematical Systems Theory, A Volume in Honor of Diedrich Hinrichsen*, pages 135–150. Birkhäuser, Boston, 2001.

[5] Jørn Justesen and Søren Forchhammer. *Two-Dimensional Information Theory and Coding*. Cambridge University Press, Cambridge, UK, 2010.

[6] Diego Napp, Carmen Perea, and Raquel Pinto. Input-state-output representations and constructions of finite support 2D convolutional codes. *Advances in Mathematics of Communications*, 4(4):533–545, 2010.

[7] Theo Pavlidis, Jerome Swartz, and Ynjiun P. Wang. Information encoding with two-dimensional bar codes. *Computer*, 25(6):18–28, 1992.

[8] Jaswinder Singh and Maninder Lal Singh. A new family of two-dimensional codes for optical CDMA systems. *Optik - International Journal for Light and Electron Optics*, 120:959–962, 2009.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On some properties of color morphology operators

Pedro Huidobro[1], Pedro Alonso[1], Susana Montes[2] and Irene Díaz[3]

[1] *Department of Mathematics, University of Oviedo, Spain*

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

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

emails: UO231995@uniovi.es, palonso@uniovi.es, montes@uniovi.es, sirene@uniovi.es

Abstract

Analysis of color images still represents a challenge due to the difficulties associated to working in high order spaces. Among the different techniques for image processing and analysis, Mathematical Morphology comprises powerful non linear techniques for filtering, texture analysis, shape analysis, edge detection or segmentation for black and white or gray scale images. The extension of Mathematical Morphology operators to multi-valued functions, and in particular to color images, is neither direct nor general due to the vectorial nature of the data. In addition, one of the problems arising when applying Mathematical Morphology is the false color problem. To avoid it, some invariance properties for these operators play an important role. In this paper, these properties are studied associated to basic morphological operators for color images.

Key words: Color Images, Morphological Operators, Erosion, Dilation, Duality, Extensitivity, Separability.

1 Introduction

Techniques of Artificial Vision have been initially developed for binary and gray scale images, where the information is codified by 2 and 2^{n+1} with $n \in \mathbb{N}$ levels, respectively. Nevertheless, the color is an important source of information. For this reason, during the last years these techniques have being developed for color images. However, nowadays, the representation and the treatment of color images are still open problems [1, 2].

 ISBN: 978-84-608-6082-2

Mathematical Morphology (MM) is the natural area for studying many problems in image analysis such as filtering, texture analysis, shape analysis, edge detection or segmentation. In the eighties, Matheron and Serra [9, 10, 12] proposed the last mathematical formulation of morphology within the algebraic framework of the lattices. The extension of MM to gray scale images is got through Fuzzy Mathematical Morphology (FMM), which is based on redefining the set operations as fuzzy set operations [3, 4, 5, 6]. Following the same strategy MM can be extended to color image spaces by defining morphological operators over the color spaces. The performance of these operators depends on the properties they satisfy. The goal of this paper is to study the properties satisfied by a certain class of morphological operators for color image processing.

This paper is organized as follows: Section 2 presents the general concepts of the MM for color images; Section 3 describes the properties satisfied by the afore defined operators. Finally, the last section details some conclusions and discusses possible future lines of research.

2 Mathematical Morphology for color images

MM is a non linear theory for spatial analysis of images where the topological relations and the geometry of the objects in the image are the parameters characterizing the object under study [7, 9, 11, 12]. The main idea of this methodology is the decomposition of a operator into a combination of the basic operators: erosion, dilation, anti-erosion and anti-dilation, as well as the supreme and the infimum operations.

Color Mathematical Morphology (CMM) can be developed from the existing theory of MM for gray levels images. In that case, it becomes necessary the definition of a complete lattice in the color space representing the chromatic information of the digital images. However, there is not a natural order for multidimensional data and therefore the extension of MM to CMM is not straightforward.

The vectorial processing of color images is based on applying a unique operation to the image considering it as an indivisible composition of vectorial pixels (see [8]). In such approach the notion of a complete lattice arises and therefore the definition of a total order over the subset τ of $\mathbb{R}^3$ becomes necessary. Since there is not any natural order in these sets, it is necessary to establish an appropriate order on color space τ. In addition, it is necessary to introduce the concept of a Structuring Elements (SE).

2.1 Structuring Element

MM examines the geometrical structures of the image by checking a small neighborhood, called SE, in different parts of the image. The SE is a completely defined set whose geometry is known in advance. It is compared to the image through translations. The size and shape of the SEs are chosen a priori depending on the morphology of the set over which it interacts

and also according to the desired shape extraction. The SEs is then moved, so that it covers the whole image pixel by pixel, making a comparison between each element and the image.

To define the operators of the CMM, the SE plays an important role. In this case, the SE is not a set as in the binary case or a function as in the case of gray levels images, but it indicates the neighborhood over which the pixels of the image are compared.

Definition 1 *Let $f : D_f \subset \mathbb{R}^2 \to \mathbb{R}^3$ be a color image, let $x \in D_f$ be a pixel, and let (D_f, d) be a metric space. A* **Structuring Element** *(SE for short) for a color pixel x is a neighborhood:*

$$B(x, r) = \{y \in D_f : d(x, y) \leq r\} \tag{1}$$

where r denotes any positive real number, which is called **diameter** *and d is a metric or distance function on D_f.*

Notice that the function d is a metric or distance function on D_f, that is, for any $x, y, z \in D_f$, d satisfies the following properties:

i. $d(x, y) \geq 0$,

ii. $d(x, y) = 0 \Leftrightarrow x = y$,

iii. $d(x, y) = d(y, x)$,

iv. $d(x, y) \leq d(x, z) + d(z, y)$.

2.2 Erosion and Dilation for color images

As it was previously mentioned, the definition of the basic operators requires a complete lattice in the color image space. This complete lattice depends on an order between images, based on an order on $\mathbf{R}^3$. It is defined as follows:

Proposition 1 *([12]) Let $\leq_\tau$ be an order on $\tau \subset \mathbf{R}^3$. The space of functions from D_f to τ with the order $\leq$ defined as*

$$f \leq g \Leftrightarrow \ f(x) \leq_\tau g(x), \qquad \forall x \in D_f$$

for any $f, g : D_f \subseteq \mathbb{R}^2 \to \tau \subseteq \mathbb{R}^3$ has a lattice structure.

The basic operators for color images are now introduced.

 ISBN: 978-84-608-6082-2

Definition 2 *Let $\tau \subset \mathbb{R}^3$ be a color space with a structure of complete lattice provided by a total order $\leq_\tau$. Let B be a SE, the basic operators* **erosion** *($\epsilon_B^{\leq_\tau}$) and* **dilation** *($\delta_B^{\leq_\tau}$) associated to a color image f are defined as follows:*

$$\epsilon_B^{\leq_\tau}(f) = \inf_{s \in B}\{f \circ T_x\} = \inf_{s \in B}\{f(x+s)\} \tag{2}$$

$$\delta_B^{\leq_\tau}(f) = \sup_{x \in B}\{f \circ T_{-x}\} = \sup_{s \in B}\{f(x-s)\} \tag{3}$$

being $T_x : \mathbb{R}^2 \to \mathbb{R}^2$ the translation function by the element $x \in \mathbb{R}^2$, that is, $T_x(s) = s + x$.

Example 1 *Let f be a color image in RGB space, B the SE defined by d_∞ ($d_\infty(x,y) = \max_{1 \leq i \leq 2}\{|x_i - y_i|\}$) with $r = 3$ and x the central pixel of the SE. Figure 1 shows the different objects f, B, x and the decomposition of the neighborhood defined by B in the three components of the RGB space. As an example, we take some values in each component.*

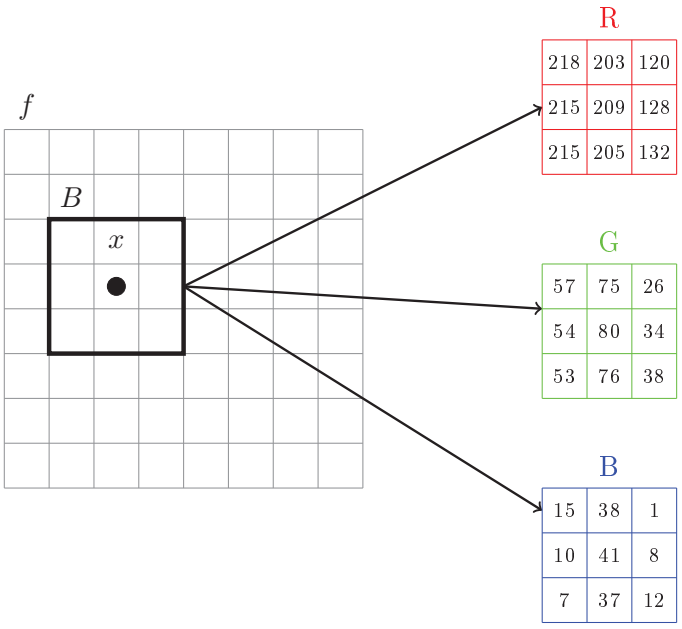


Figure 1: Image decomposition in a neighborhood.

Therefore, to define the erosion or dilation is necessary to sort the pixels of the neighborhood to determine the infimum or maximum respectively. The pixels to sort are the following: $x_1 = (218, 57, 15)$; $x_2 = (215, 54, 10)$; $x_3 = (215, 53, 7)$; $x_4 = (203, 75, 38)$; ...; $x_9 = (132, 38, 12)$ (see Figure 1).

To obtain the infimum and maximum it is neccessary to define an order to sort the pixels.

From the combination of these basic operators, erosion and dilation, other operations can be defined in the same way they were defined for binary and gray levels images [7, 12], as we can see in the following definition.

Definition 3 *Let f a color image and B be a SE, the following morphological operators are defined as:*

- **Morphological Gradient**: $Grad_B^{\leq_\tau}(f) = \delta_B^{\leq_\tau}(f) - \varepsilon_B^{\leq_\tau}(f)$
- **Gradient by erosion**: $Grad\_Ero_B^{\leq_\tau}(f) = f - \varepsilon_B^{\leq_\tau}(f)$
- **Gradient by dilation**: $Grad\_Dil_B^{\leq_\tau}(f) = \delta_B^{\leq_\tau}(f) - f$
- **Opening**: $\gamma_B^{\leq_\tau}(f) = \delta_B^{\leq_\tau}(\varepsilon_B^{\leq_\tau}(f))$
- **Closing**: $\phi_B^{\leq_\tau}(f) = \varepsilon_B^{\leq_\tau}(\delta_B^{\leq_\tau}(f))$

Let us remark that the erosion and dilation of a color image strongly depend on the order established in the color space. Therefore, it has not a unique meaning as in the case of the operators of the MM for binary and gray levels images. Thus, the performance of the previous defined operators depend on the order established.

3 Properties of Operators

As stated in [13] a morphological operator must satisfy certain properties. In this section, it is checked if these properties are satisfied by the previous operators and thus they can be considered as morphological operators. In fact, we analyze the increasing (also called order-preserving), the extensive (the input always less than or equal to the output) and idempotent (repeated application has no effect), among others.

Proposition 2 *Previous erosion (ε) and dilation (δ) operators satisfy the following properties, being f and g color images:*

1 ***Duality***

$$(\varepsilon(f))^c = \delta(f^c),$$

with f^c being the complementary image of f defined as $f^c = (1,1,1) - f$.

 ISBN: 978-84-608-6082-2

2 ***Order-preserving***.

$$f \leq g \Rightarrow \begin{cases} \varepsilon(f) \leq \varepsilon(g), \\ \delta(f) \leq \delta(g). \end{cases}$$

3 ***Extensitivity***: *If the SE contains the origin, dilation is extensive and erosion anti-extensive, i.e.,*

$$0 \in B \Rightarrow \begin{cases} \varepsilon(f) \leq f, \\ f \leq \delta(f). \end{cases}$$

4 ***Separability***: *SE can be divided into pieces. Erosion and dilation can be applied by unidimensional erosions and dilations*

$$B = \delta_{B_1}(B_2) \Rightarrow \begin{cases} \varepsilon_B(f) = \varepsilon_{B_1}(\varepsilon_{B_2}(f)), \\ \delta_B(f) = \delta_{B_1}(\delta_{B_2}(f)). \end{cases}$$

Regarding to the opening and closing, the next properties are satisfied.

Proposition 3 *Opening and closing operators satisfy the following properties:*

1 ***Increasing***: *It is straightforward as opening and closure are composition of increasing functions.*

2 ***Extensitivity***: *Opening is anti-extensive, closing is extensive.*

3 ***Idempotence***: *Both, opening and closing are idempotent.*

4 Conclusions

This paper revises the properties satisfied by a new family of morphological operators for color images. These properties are required in order to avoid some of the most important problems related to color image processing, such as the false color problem. In future works we will try study some invariance properties allowing us to build invariant color representations to obtain the best performance in several tasks.

Acknowledgements

The research in this paper has been supported in part by grant MINECO-TIC2014- 59543-P; its financial support is gratefully acknowledged.

References

[1] J. Angulo, *Morphological color operators in totally ordered lattices based on distances: Application to image filtering, enhancement and analysis*, Computer Vision and Image **107(1)** (2007) 193–204.

[2] E. Aptoula, S. Lefevre, *On lexicographical ordering in multivariate mathematical morphology*, Pattern Recognition Letters **29(2)** (2008) 109–118.

[3] B. De Baets, *Idempotent closing and opening operations in fuzzy mathematical morphology*, In Uncertainty Modeling and Analysis, 1995, and Annual Conference of the North American Fuzzy Information Processing Society, Proceedings of ISUMA-NAFIPS'95, pages 228–233.

[4] I. Bloch, *Duality vs. adjunction for fuzzy mathematical morphology and general form of fuzzy erosions and dilations*, Fuzzy Sets and Systems **160(13)** (2009) 1858–1867.

[5] I. Bloch, H. Maître, *Fuzzy mathematical morphologies: a comparative study*, Pattern recognition **28(9)** (1995) 1341–1387.

[6] P. Burillo López, N. Frago Pa nos, R. Fuentes-González, *Fuzzy morphological operators in image processing*, Mathware & Soft Computing **10(2)** (2003) 85–100.

[7] R. González, R. Woods, *Digital Image Processing*, Addison-Wesely Publishing Company, 1996.

[8] P. Lambert, J. Chanussot, *Extending mathematical morphology to color image processing*, In CGIP'00-1st International Conference on Color in Graphics and Image, pages 158–163, 2000.

[9] G. Matheron, *Random sets and integral geometry*, John Wiley & Sons Inc, New York, 1975.

[10] C. Ronse, *Why mathematical morphology needs complete lattices*, Signal processing **21(2)** (1990) 129–154.

[11] J. Serra, *Image analysis and mathematical morphology (Vol. I)*, Academic Press, London, 1982.

[12] J. Serra, *Image analysis and mathematical morphology (Vol. II)*, Academic Press, London, 1988.

[13] J.J. van de Gronde, J.B. Roerdink, *Group-invariant color morphology based on frames*, IEEE Transactions on Image Processing **23(3)** (2014) 1276–1288.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Tensor Rank Decomposition of the Coulomb Integrals

Felix Hummel[1] **and Andreas Grüneis**[1]

[1] *Solid State Research, Max-Planck Institute, Stuttgart, Germany*

emails: f.hummel@fkf.mpg.de, a.grueneis@fkf.mpg.de

Abstract

Coupled Cluster methods employ the Coulomb integrals, a tensor of fourth order, for approximating the electronic correlation energy. Its memory footprint scales like $\mathcal{O}(N^4)$, which often poses the bottleneck of Coupled Cluster methods.

We study a decomposition of the Coulomb integrals into a tensor contraction of six matrices of which only two are distinct. We find that the Coulomb integrals of insulators as well as of metals can be well approximated in this form already with small matrices compared to the number of real space grid points. The matrices can be computed in $\mathcal{O}(N^4)$ using a regularized form of the Alternating Least Squares (ALS) algorithm.

The studied factorization of the Coulomb integrals can be exploited to reduce the computational costs of the Distinguishable Cluster Singles Doubles (DCSD) method to $\mathcal{O}(N^5)$.

Key words: tensor rank decomposition canonical-polyadic-decomposition cpd

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Insights into the Bonding Situation of Interstitial Gold Clusters and Ligand Stabilized Au(0) Complexes

Paul Jerabek[1] and Gernot Frenking[2]

[1] *Centre for Theoretical Chemistry and Physics (CTCP), The New Zealand Institute for Advanced Study (NZIAS), Massey University, Auckland, 0745 Auckland, New Zealand*

[2] *FB Chemie, Philipps-Universität Marburg, 35037 Marburg, Germany*

emails: P.Jerabek@massey.ac.nz,

Abstract

Experimentally known gold compounds were analyzed with respect to their bonding situation by the means of the density functional theory (DFT). Several quantum chemical tools were used for this investigation: The Quantum Theory of Atoms in Molecule (QTAIM), Energy Decomposition Analysis with Natural Orbitals for Chemical Valence (EDA-NOCV) and the analysis of Frontier Molecular Orbitals (FMO).
The bonding situation of the icosahedral gold cage compounds $[M(AuPH_3)_{11}AuCl]^{3+}$ (M = Pt, Pd, Ni) is a cross between a metallic cluster and complex-like interactions of the ligands with the central atom.
An unusual bonding situation is present in the Au(0) complex (cAAC)→Au←(cAAC) (cAAC = Cyclic (alkyl) (amino) carbene): The gold atom possesses an unpaired electron which is delocalized over the p(π) AOs of the carbene ligands and the p orbital of Au. Through EDA-NOCV analyses, the strongest interaction between the Au and (cAAC) ligands could be identified as the donation of the unpaired p electron of the Au to the (cAAC) ligands, meaning that gold is acting as a strong donor in this system.

Key words: Gold, Au(0), Interstitial Clusters, Complexes, Bonding Analysis, Energy Decomposition Analysis (EDA-NOCV)

1 Introduction

Some transition metal clusters can not be synthesized without the stabilizing effect of an interstitial atom in the center of the cage.[1, 2] Besides the attractive interactions of the metal atoms within the cage, the interactions between the enclosed central atom and the

metallic cage have to be considered as well. In such cases, the bonding situation is not easily describable with a single model and the systems need to be viewed as metallic clusters and transition metal complex compounds simultaneously. The experimentally accessible compounds $[M(AuPH_3)_{11}AuCl]^{3+}$ (M = Pt, Pd, Ni) are examples for such systems.[3] They fulfill the 18 valence electron (VE) rule for stable transition metal complexes, but at the same time can be understood as stable gold clusters with 8 cluster valence electrons (CVE) in a $1S^21P^6$ superatom configuration.[4]

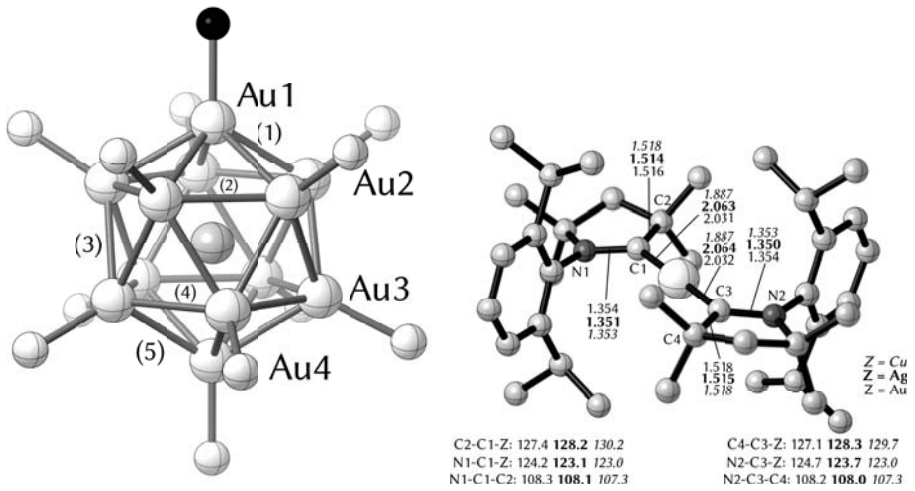


Figure 1: *Left:* General structure of the interstitial gold cage compounds $[M(AuPH_3)_{11}AuCl]^{3+}$. *Right:* M06/SVP optimized structures of (cAAC)→Au/Ag/Cu←(cAAC).

On the other hand, it is possible to reduce the number of cluster atoms and arrive at the ligand stabilized complex (cAAC)→Au←(cAAC) (cAAC = Cyclic (alkyl) (amino) carbene).[5, 6] This system has the gold atom in an oxidation state of 0 and can therefore be described as the "smallest piece of gold". A comparison with the homologues (cAAC)→Cu←(cAAC)[7] and the hypothetical (cAAC)→Ag←(cAAC) shows the significance of relativistic effects for the stability of the gold complex.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG).

References

[1] P. Pyykkö, N. Runeberg, *Angew. Chem. (Int. Ed.)* **2002**, *41*(12), 2174–2176.

[2] X. Li, B. Kiran, J. Li, H. J. Zhai, L. S. Wang, *Angew. Chem. (Int. Ed.)* **2002**, *41*(24), 4786–4789.

[3] A. Puls, P. Jerabek, W. Kurashige, M. Molon, T. Bollermann, M. Winter, C. Gemel, Y. Negishi, G. Frenking, R. A. Fischer, *Angew. Chem. (Int. Ed.)* **2014**, *53*, 4327–4331.

[4] H. Häkkinen, *Chem. Soc. Rev.* **2008**, *37*(9), 1847–1859.

[5] P. Jerabek, H. W. Roesky, G. Bertrand, G. Frenking, *J. Am. Chem. Soc.* **2014**, *136*, 17123–17135.

[6] D. S. Weinberger, M. Melaimi, C. E. Moore, A. L. Rheingold, G. Frenking, P. Jerabek, G. Bertrand, *Angew. Chem. (Int. Ed.)* **2013**, *52*(34), 8964–8967.

[7] D. S. Weinberger, N. Amin SK, K. C. Mondal, M. Melaimi, G. Bertrand, A. C. Stückl, H. W. Roesky, B. Dittrich, S. Demeshko, B. Schwederski, W. Kaim, P. Jerabek, G. Frenking, *J. Am. Chem. Soc.* **2014**, *136*(17), 6235–6238.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Efficient Implementation of Morphological Index for Building/Shadow Extraction from Remotely Sensed Images

Luis Ignacio Jiménez[1], Javier Plaza[1] and Antonio Plaza[1]

[1] *Hyperspectral Computing Laboratory, Department of Computer Technology and Communications, University of Extremadura*

emails: luijimenez@unex.es, jplaza@unex.es, aplaza@unex.es

Abstract

Morphological building index (MBI) and morphological shadow index (MSI) are recently developed techniques that aim at automatically detect buildings/shadows using high resolution remotely sensed imagery. Traditional mathematical morphology operations are usually time-consuming as they are based in the consideration of a wide range of image-object properties such as brightness, contrast, shapes, sizes and in the application of series of repeated transformations (e.g. classical opening and closing operators). In the case of MBI and MSI, the computational complexity is also increased due to the use of multiscale and multidirectional morphological operators. In this paper, we provide a computationally efficient implementation of MBI and MSI algorithms which is specifically developed for commodity graphic processing units (GPUs) using Nvidia CUDA. We perform the evaluation of the parallel version of the algorithms using two different NVidia architectures and three widely used hyperspectral datasets. Experimental results show that the computational burden introduced when considering multidirectional morphological operators can be almost completely removed by the developed implementations.

Key words: mathematical morphology, high resolution, remotely sensed imagery, graphic processing units (GPUs).

1 Introduction

The efficient and precise location/identification of buildings is an increasingly important task for a great part of the most developed countries of the world, as it provides crucial information for population estimation and territorial planning [1]. This is possible due to the

availability of high-resolution Earth Observation (EO) instruments that now provide almost complete spatial information about the surface of the Earth that can be efficiently used to complement available spectral information [2]. This allows for an increase in the separability of spectrally similar classes. With this purpose, several sophisticated (supervised or semi-supervised) segmentation techniques have been developed for building extraction [3, 4, 5].

Recently, most efforts have been focused on the generation of a feature index that can be applied to building detection without the need for training data or complex segmentation processes. The morphological building/shadow index has been recently proposed with this aim [6]. The main idea of MBI is to relate the implicit characteristics of buildings (e.g. brightness, size and contrast) with morphological operators (e.g. top-hat reconstruction, granulometry and directionality).

As mentioned before, while integrated spatial/spectral developments hold great promise for Earth science image analysis, they clearly introduce new processing challenges that, combined with the complex and large size of EO datasets, limits the possibility of utilizing those algorithms in time-critical applications [7, 8]. Particularly, the use of multiscale and multidirectional morphological operators introduces a significant computational burden in MBI algorithm [9] which can be alleviated if parallel implementations are developed.

Even though EO data processing algorithms map nicely to clusters and heterogeneous networks [10, 11], these systems are generally expensive and difficult to adapt to on-board data processing scenarios, in which low-weight and low-power integrated components are essential to reduce mission payload where field programmable gate arrays (FPGAs) and graphic processing units (GPUs) can further provide a response in (near) real time, which is believed to be acceptable in many remote sensing applications [12]. In this paper we present the first GPU-based parallel implementation of the MBI/MSI algorithm for EO data exploitation using the NVidia CUDA framework.

The remainder of the paper is organized as follows. Section II describes the original implementation of MBI and MSI algorithms and our proposed C implementation (optimized for memory usage). Section III briefly introduces GPU architectures and the NVidia CUDA framework, and further describes the newly proposed GPU implementation for MBI and MSI algorithms. Section IV evaluates the proposed GPU implementations in terms of building/shadow detection accuracy and computational performance. Section V concludes this paper with some remarks and hints at plausible future research lines.

2 Morphological Building/Shadow Index

Morphological building/shadow index is based on the construction of a relationship between the implicit characteristics of buildings (e.g. brightness, size and contrast) with morphological operators (e.g. top-hat reconstruction, granulometry and directionality). Some of the basics of this relationship are introduced below.

- Brightness. We can define the brightness of a pixel as the maximum value of the pixel at all the contained spectral bands. Building areas are characterized by high brightness scores while shadow areas brightness should be smaller due to their low spectral reflectance. White/black top-hat transformation will be used to point out bright/dark structures with a determined size in order to identify buildings/shadows.
- Contrast. Building areas are generally characterized by their high contrast, due to the difference between the spectral reflectance values of roof and spatially adjacent shadows. The case of shadows is exactly the opposite (high contrast between shadows and the neighboring areas) The MBI algorithm is able to characterize the local contrast of buildings extracting the differential morphological profiles (DMP) of the white top-hat transformed data. MSI algorithm relies on the extraction of the DMP over the black top-hat transformed data.
- Shape. The most widely adopted shape descriptor for building areas is the rectangle. Therefore, the length-width ratio can be used to filter out structures with similar spectral response.
- Size and directionality. In order to assist with the removal of spectrally similar structures, a series of linear structuring elements (SEs), designed to measure the size and directionality of structures, is implemented in both MBI and MSI.

Based on the above concepts, we can describe the steps of the MBI algorithm as follows (the main differences betwee MSI and MBI are also highlighted):

1. Calculation of the brightness. Using all the spectral bands of the considered image, we select the brightest component of each pixel as follows:

$$b(x) = \max_{1<=k<=K} band_k(x), \tag{1}$$

where x is the pixel, K is the number of components of the pixel's spectral signature, and $band_k(x)$ is the pixel value in the k-th band.

2. The white top-hat by reconstruction for MBI and the black top-hat by reconstruction for MSI (2). White top-hat is then computed to highlight bright structures:

$$W - TH(d, s) = b - \gamma_b^{re}(d, s), \tag{2}$$

while the black top-hat aims at highlighting dark structures:

$$B - TH(d, s) = \varphi_b^{re}(d, s) - b, \tag{3}$$

where γ_b^{re} represents the opening by reconstruction, φ_b^{re} represents the closing by reconstruction, and d denotes the size and directionality of the linear SE.

 ISBN: 978-84-608-6082-2

3. The morphological profiles (MPs) of the white top-hat are now defined as:

$$\begin{cases} MP_{W-TH}(d,s) = W - TH(d,s) \\ MP_{W-TH}(d,0) = b \end{cases} \tag{4}$$

4. To complete the calculation of the MBI and MSI we need to define the differential morphological profiles (DMP) as:

$$DMP_{W-TH}(d,s) = |MP_{W-TH}(d,s+\Delta s) - MP_{W-TH}(d,s)|, \tag{5}$$

$$DMP_{B-TH}(d,s) = |MP_{B-TH}(d,s+\Delta s) - MP_{B-TH}(d,s)|, \tag{6}$$

where Δs is the interval of the profiles between and $s_{min} <= s <= s_{max}$.

5. The MBI and MSI are defined as the average of the DMP of the white top-hat and the black top-hat, respectively:

$$MBI = \frac{\sum DMP_{W-TH}(d,s)}{D * S}, \tag{7}$$

$$MSI = \frac{\sum DMP_{B-TH}(d,s)}{D * S}, \tag{8}$$

where D is the number of the directions applied to the linear SE and $S = ((s_{max} - s_{min})/\Delta s) + 1$. Buildinds and shadows are respectively represented by larger values in each index.

Algorithm 1 provides a pseudode description of the original MBI/MSI algorithm implemented in Matlab. When calculating the black top-hat and white top-hat, it should be noticed that opening and closing by reconstruction morphological operations are complementary.

For the C version we introduce some changes in order to optimize the code to the new language (specially focusing on the memory management). In the Matlab version, iteration t partially repeats operations from the previous iteration $(t-1)$ oriented to the white top-hat and black top-hat creation, while in the C implementation, we preserve the operations already executed in previous iterations due to the absence of memory restrictions. Besides, we eliminate the creation of the linear SE by applying the direction of it to the erosion morphological operator. The resulting pseudocode of the C implementation is provded in Algorithms 2 and 3. Using this latest version as a reference, we have developed a CUDA version that implements each step as a CUDA kernel. This version is described in the following section.

 ISBN: 978-84-608-6082-2

Algorithm 1 Pseudocode of **Morphologic Building Index** and **Morphologic Shadow Index** in Matlab code

```
image = ReadImage
img = CalculationBrightest/DarkestScene(image)
for s = smin:Δs:smax
   for dir = 1:1:D
         se = CreateStructureElement(dir)
         a = W-TH(se,s,img) / B-TH(se,s,imgC)
         b = W-TH(se,s+Δs,img) / B-TH(se,s+Δs,imgC)
         DMP(dir) = b - a
   end
end
DMP = DMP/(D*S)
MBI/MSI = ∑DMP(dir)
Scale(MBI,0,1)
```

3 Parallel Implementation

GPUs can be understood in terms of a stream model, under which all data sets are represented as streams (i.e., ordered data sets), and each of them is processed by a multiprocessor, which means that a GPU also can be seen as a set of multiprocessors (MPs). Each multiprocessor is characterized by a single instruction multiple data (SIMD) architecture. Each processor has access to a local shared memory and also to local cache memories in the multiprocessor, while the multiprocessors have access to the global GPU (device)memory. Algorithms are constructed by chaining so-called *kernels* which operate on entire streams and are executed by a multiprocessor, taking one or more streams as inputs and producing one or more streams as outputs. The kernels can perform a kind of batch processing arranged in the form of a *grid* of *blocks*, where each block is composed by several *threads* which share data efficiently through the shared local memory and synchronize their execution for coordinating accesses to memory. As a result, there are different levels of memory in the GPU for the thread, block and grid concepts. There is also a maximum number of threads that a block can contain but the number of threads that can be concurrently executed is much larger (several blocks executed by the same kernel can be managed concurrently, at the expense of reducing the cooperation between threads since the threads in different blocks of the same grid cannot synchronize with the other threads). Our GPU implementation of MBI is based in the following kernels.

BrightnessImage: this kernel implements the first step of MBI and MSI algorithm, where the brightness is calculated according to equation 1. In the case of MSI, a final step to complement the returned structure is performed. The number of threads is set to

 ISBN: 978-84-608-6082-2

Algorithm 2 Pseudocode of **Morphologic Building Index** in C code

```
image = ReadImage
img = CalculationBrightestScene(image)
for s = smin:Δs:smax
   for dir = 1:1:D
         erode=Erosion(dir, img)
         recon=Reconstruction(erode,img)
         wth1(dir) = img - recon
         if(s != smin)
            DMPWTH(dir) += (wth2(dir)-wth1(dir))
         else
            DMPWTH(dir) = 0
         wth2(dir) = wth1(dir)
   end
end
MBI = ∑DMPWTH(dir)
Scale(MBI,0,1)
```

the maximum allowed by the device and the number of blocks equals the number of pixels divided by the number of threads.

ErodeOperator: the morphological erosion operator is implemented using different kernels depending on the direction of the linear SE being applied. The original work explains that the number of directions is set to four (NE, N, NW and W) because changing to an eight-connected neighborhood resulted in a similar outcome with a significant computational time increase. As result, one kernel computes the erosion for the four directions storing the erode images consecutively in memory. This kernel uses a two-dimensional grid setting the x-dimension to the number of lines and the y-dimension to the number of samples both divided by the block size for each dimension, which is the same making a square block of size 32.

Reconstruction: this step performs the morphological reconstruction using two kernels that performs the raster scan (*x_forward*) and the antiraster scan (*x_backward*) of the four directions erode images; *x_forward* finds the maximum value within the NE, N, NW, W neighbors and the origin pixel from the top-left to the bottom-right of the image; and *x_backward* computes the maximum value within the E, SE, S, SW neighbors and the origin pixel from the bottom-right to the top-left of the image. The same way as the erosion kernel, the reconstruction is performed for the four directions considered in the same call. Both kernels set the number of block to the number of samples divided by a number of threads empirically set to 32, in order to maintain a balance between blocks and threads.

 ISBN: 978-84-608-6082-2

Algorithm 3 Pseudocode of **Morphologic Shadow Index** in C code

```
image = ReadImage
img = CalculationDarkestScene(image)
for s = smin:Δs:smax
   for dir = 1:1:D
         erode=Erosion(dir, img)
         recon=Reconstruction(erode,img)
         wth1(dir) = recon - img
         if(s != smin)
            DMPWTH(dir) += (wth2(dir)-wth1(dir))
         else
            DMPWTH(dir) = 0
         wth2(dir) = wth1(dir)
   end
end
MBI = ∑DMPWTH(dir)
Scale(MSI,0,1)
```

Substraction: this stage computes, in one kernel called *substract*, the difference between consecutive iterations reconstructed images, acumulating the results to perform the average in a subsequent step. Once the iterave process is finished, other kernel performs the averaging of the results based in the number of iterations between the minimum,*smin*, and maximum, *smax*, structure size values used.

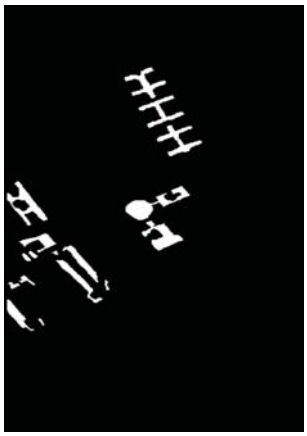
4 Experimental validation

4.1 Hyperspectral data and hardware architectures

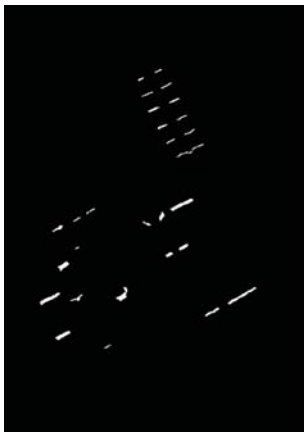
Our experiments have been carried out using three different hyperspectral images. The first considered hyperspectral image is the well-known Pavia University hyperspectral dataset (Figure 1a), acquired by the ROSIS optical sensor during a flight campaign over the urban area of the University of Pavia, Pavia, Italy. The original Pavia University dataset consists on $610 \times 340 pixels$, with high spatial resolution of 1.3 m per pixel. The number of data channels in the acquired image is 103 (with the spectral range from 0.43 to 0.86 μm). Nine thematic land-cover classes are available, from which we select metal sheets, self-blocking bricks and bitumen to generate the class building (see Figure 1b). In addition, a shadow class is also provided in the ground-truth information (see Figure 1c).

The second hyperspectral dataset used was acquired by the NSF-funded Center for Airborne Laser Mapping (NCALM) over the University of Houston campus and its neighboring

(a) Pavia University image



(b) Building ground-truth



(c) Shadow ground-truth

Figure 1: Pavia University hyperspectral dataset. (a) False color composition of the Pavia University image. (b) Reference spatial distribution of the buildings. (c) Reference spatial distribution of the shadow class.

area. This hyperspectral data has 144 spectral bands in the 380-1050 nm spectral region and spatial resolution of 2.5 m. The image size in pixels is 349 $\times$ 1905. Figure 2a shows a false color composite of the image. Ground-truth information is available as 15 different land-cover classes. The building groud-truth has been generated by the fusion of residential and commercial original land-cover classes.

The last hyperspectral image scene used for experiments in this work was collected by the AVIRIS sensor, which was flown by NASAs Jet Propulsion Laboratory over the World Trade Center area in New York City on 16 September 2001, just 5 days after the terrorist attacks that collapsed the two main towers and other buildings in the WTC complex. The selected subset consists of 500 $\times$ 1600$pixels$, 224 spectral bands, and a total size of (approximately). The spatial resolution is 1.7 m/pixel. Extensive reference information, collected by U.S. Geological Survey (USGS), is available for the WTC scene.

We have used two different computer architectures for the experimental validation of the proposed approaches: a compute cluster with 44 NVidia TESLA S2070 GPU nodes (2 M2075 per node), each with an Intel Xeon CPU E5645 at 2.40 GHz and a total of 24 GB of RAM, divided in 12 modules of 2 GB each (hereinafter Architecture 1) and a desktop computer (Intel i7 920 CPU at 2.67 GHz and 6 GB of RAM) with an NVidia GTX 580 GPU equipped with 512 processor cores operating at 1.54 GHz and 1536 MB of RAM memory (hereinafter Architecture 2).

[1]http://www.ceta-ciemat.es

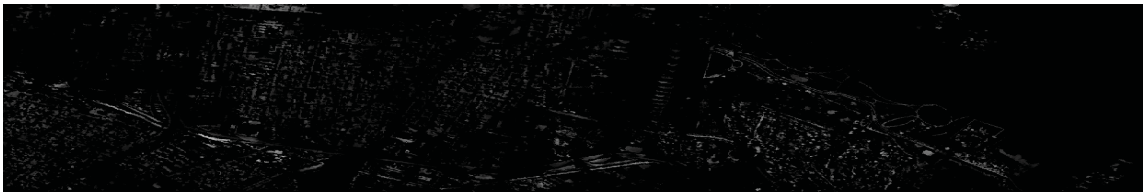
(a) Houston campus image

(b) Building ground-truth

Figure 2: Houston campus hyperspectral dataset. (a) False color composition of the Houston campus image. (b) Reference spatial distribution of the buildings.

(a) World Trade Center image

(b) Building Index



(c) Shadow Index

Figure 3: False color composition of the World Trade Center hyperspectral dataset acquired by the AVIRIS instrument.

 ISBN: 978-84-608-6082-2

4.2 Analysis of algorithm precision

In this section we will focus on analyzing the parallel MBI and MSI implementations using the two datasets with groud-truth information about building and shadows. Particularly, Table 1 shows results based on the generation of a binary image applying different threshold values ($th1 = 1/255$, $th2 = 50/255$, $th3 = 100/255$, $th4 = 150/255$ and $th5 = 200/255$) over the MBI and MSI estimation images over Pavia University and Houston datasets (we remove all the pixels with a value below the thereshold, considering the range of 0 to 255 based on the original RGB algorithms scale). Then, we calculate the MBI/MSI values (being 100 the best case and 0 the worst) by comparing the each thresholded image with the ground-truth. It can be seen that the results obtained by the CPU and GPU implementations are almost the same (the slight differences are due to the removal of the queue structure, which seems to benefit the parallel implementations). It should be noticed that no ground-truth information is available for the shadow class in the Houston dataset and, therefore, no precision results can be shown for this particular case.

		Pavia Univ. (610×340)					Houston Univ. (349×1905)				
Algorithm	Implementation	th1	th2	th3	th4	th5	th1	th2	th3	th4	th5
MBI	CPU	100	18,79	10,68	2.61	0	100	5.85	4.10	1.31	0
	GPU	100	18,24	10,45	2,84	0	100	6,68	4.57	1,79	0
MSI	CPU	100	97.36	47.94	19.64	3.69	–	–	–	–	–
	GPU	100	97,99	50,05	18,80	6.44	–	–	–	–	–

Table 1: Mean execution time (in seconds) for the CPU and GPU implementations along with the obtained speedup after 10 Monte-Carlo runs over each of the considered architectures over the two processed hyperspectral datasets with available groundtruth.

4.3 Analysis of parallel performance

Table 2 shows the obtained speedups in the two considered architectures for the three selected scenes. The results obtained by the Architecture 2 are better due to the fact that the NVidia TESLA S2070 includes error checking and correction (NVidia GTX 580 does not include this characteristic) that guarantees more stable results at the expense of a slightly reduced performance. The time taken by data transfers between the CPU and the GPU is included in the execution times. As it can be seen, speedups around $\times 5$ can be achieved when considering the two large datasets. It is important to emphasize that parallel implementation is able to overlap the processing of the four considered directions, thus providing a significant performance improvement with regards to the serial implementation.

		Pavia Univ. (610 × 340)			Houston Univ. (349 × 1905)			WTC (500 × 1600)		
	Hardware	CPU time	GPU time	Speedup	CPU time	GPU time	Speedup	CPU time	GPU time	Speedup
MBI	Architec. 1	0.794	0.508	x1.563	2.735	0.608	x4.498	3.850	0.886	x4.345
	Architec. 2	0.695	0.426	x1.632	2.575	0.496	x5.188	3.537	0.708	x4.995
MSI	Architec. 1	0.809	0.503	x1.608	2.792	0.605	x4.615	3.920	0.892	x4.395
	Architec. 2	0.699	0.427	x1.639	2.625	0.494	x5.314	3.556	0.704	x5.050

Table 2: Mean execution time (in seconds) for the CPU and GPU implementations of MBI and MSI along with the obtained speedup after 10 Monte-Carlo runs over each of the considered architectures over the three considered hyperspectral datasets.

5 Conclusions and future reserch lines

In this paper we have presented a GPU implementation of MBI/MSI algorithms for building/shadow detection in high resolution remote sensing images. The implementation are based in optimized implementations developed in C, which reduce the amount of memory required. In addition, an efficient raster image processing scheme is implemented on the GPU. As a result, we achieve independence between the execution time of the parallel implementation and the number of considered directions when applying the MBI/MSI multidirectional morphological operators. In our experiments, four different directions have being processed simultaneously in the GPU implementation, achieving speedups over $\times 5$ for some of the considered images. Future research lines will focus on improving both the accuracy and the computational performance of the proposed approaches. We will also explore the use of FPGAs as a specialized device with low power consumption and onboard processing capabilities in order to accelerate the MBI/MSI algorithms.

Acknowledgements

This work has been supported by Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005). This work was supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain.

References

[1] A. Marinoni and P. Gamba, "Accurate detection of anthropogenic settlements in hyperspectral images by higher order nonlinear unmixing," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 9, no. 5, pp. 952–961, 2016.

[2] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "Spatial/spectral endmember extraction by multidimensional morphological operations," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, pp. 2025–2041, 2002.

[3] M. Pesaresi and A. Gerhardinger, "Improved textural built-up presence index for automatic recognition of human settlements in arid regions with scattered vegetation," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 4, pp. 16–26, 2011.

[4] J. Plaza, A. Plaza, P. Gamba, and G. Trianni, "Efficient multi-band texture analysis for remotely sensed data interpretation in urban areas," in *Urban Remote Sensing Joint Event (IEEE URS2007), Paris, France*, 2007.

[5] F. Dell'Acqua, P. Gamba, A. Ferrari, J. A. Palmason, J. A. Benediktsson, and K. Arnason, "Exploiting spectral and spatial information in hyperspectral urban data with high resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 322–326, 2010.

[6] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral geoeye-1 imagery," *Photogramm. Eng. Remote Sens.*, vol. 77, pp. 721–732, 2011.

[7] J. Plaza, A. Plaza, and C. Barra, "Multi-channel morphological profiles for classification of hyperspectral images using support vector machines," *Sensors*, vol. 9, pp. 196–218, 2009.

[8] J. Delgado, G. Martin, J. Plaza, L. I. Jimenez, and A. Plaza, "Fast spatial preprocessing for spectral unmixing of hyperspectral data on graphics processing units," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 9, pp. 952–961, 2016.

[9] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 5, pp. 161–172, 2012.

[10] A. Plaza, J. Plaza, and H. Vegas, "Improving the performance of hyperspectral image and signal processing algorithms using parallel, distributed and specialized hardware-based systems," *signal*, vol. 50, pp. 293–315, 2010.

[11] A. Plaza, J. Plaza, and A. Paz, "Parallel heterogeneous cbir system for efficient hyperspectral image retrieval using spectral mixture analysis," *Concurrency Computat. Pract. Exper.*, vol. 22, no. 9, pp. 293–315, 2010.

[12] A. Plaza, J. Plaza, A. Paz, and S. Sanchez, "Parallel hyperspectral image and signal processing," *IEEE Signal Process. Mag.*, vol. 28, pp. 196–218, 2011.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Multi-core Implementation of Spatial-Spectral Preprocessing for Hyperspectral Unmixing

Luis Ignacio Jiménez[1], Sergio Bernabé[2], Carlos García[2], Javier Plaza[1], Gabriel Martín[3], Sergio Sánchez[1] and Antonio Plaza[1]

[1] *Department of Computer Technology and Communications, University of Extremadura*

[2] *Department of Computer Architecture and Automation, Complutense University of Madrid*

[3] *Instituto de Telecomunicacões, Lisbon*

emails: luijimenez@unex.es, sebernab@ucm.es, garsanca@ucm.es, jplaza@unex.es, gabriel.hernandez@lx.it.pt, sersanmar@unex.es, aplaza@unex.es

Abstract

Spectral unmixing pursues the identification of spectrally pure constituents, called *endmembers*, and their corresponding *abundances* in each pixel of a hyperspectral image. Most unmixing techniques have focused on the exploitation of spectral information alone. Recently, some techniques have been developed to take advantage of the complementary information provided by the spatial correlation of the pixels in the image. Computational complexity represents a major problem in these spatial-spectral techniques, as hyperspectral images contain very rich information in both the spatial and the spectral domains. In this letter, we develop a computationally efficient implementation of a spatial-spectral processing (SSPP) algorithm that has been successfully applied prior to spectral unmixing of hyperspectral data. Our implementation has been optimized for multi-core processors, and is evaluated (using both synthetic and real data) using an 2×Intel Xeon processor E5-2670 at 2.60GHz. Significant speedups can be achieved when processing hyperspectral images of different sizes. This allows for the inclusion of the proposed parallel preprocessing module in a full hyperspectral unmixing chain able to operate in real time.

Key words: Hyperspectral unmixing, spatial-spectral preprocessing (SSPP), OpenMP, multi-core processors.

 ISBN: 978-84-608-6082-2

1 Introduction

In hyperspectral unmixing, endmember extraction is the process of collecting pure signature spectra of the materials present in a remotely sensed hyperspectral scene. These pure signatures are then used to decompose the scene into a set of so-called abundance fractions representing the coverage of each endmember in each image pixel.

Several algorithms have been developed for automatic or semi-automatic identification of endmembers over the last decade [1] which majority have been developed under the pure pixel assumption, i.e., they assume that the remotely sensed data contain one pure observation for each different material in the scene [2]. Most of these algorithms rely exclusively on the exploitation of spectral information in order to select the final set of endmembers.

In order to include also the spatial information, several spatial preprocessing algorithms have been developed that can be applied prior to any spectral-based endmember extraction technique. Techniques include the spatial preprocessing (SPP) [3], region-based spatial preprocessing (RBSPP) [4], and spatial-spectral preprocessing (SSPP) [5]. The goal of these preprocessing methods is to guide the search for endmembers using not only spectral but also spatial information, which greatly assists in the selection of more spatially representative endmembers without the need to modify the endmember identification algorithm (the preprocessing can be applied as an optional step). As consequence, the spatial preprocessing increase the computational cost to the full spectral unmixing chainmaking efficient implementations for spatial preprocessing techniques an important goal.

In this work, we present a new parallel implementation of the SSPP algorithm, which has been shown as one of the most successful spatial preprocessing techniques available in the literature [5]. Our implementation has been developed for multi-core architectures where real and synthetic scenes are used to validate the efficacy of the implementation.

The remainder of this manuscript is organized as follows. Section 2 enumerates and describes the different steps of the SSPP method. Section 3 describes the proposed parallel implementation for multi-core processors. Section 4 describes the experiments conducted using real and synthetic data sets intended to evaluate the acceleration achieved by our parallel implementation. Section 5 concludes the paper with some remarks and hints at plausible future research lines.

2 Spatial-Spectral Preprocessing

This section briefly outlines the SSPP algorithm in [5]. As shown in the flowchart given in Fig. 1, the SSPP method consists of the following steps:

Multi-scale Gaussian filtering. This step takes as input the original hyperspectral image $\mathbf{Y}$ and returns a filtered version of the image. To perform this step, we first apply Gaussian filtering to each of the B spectral bands of the hyperspectral image. This

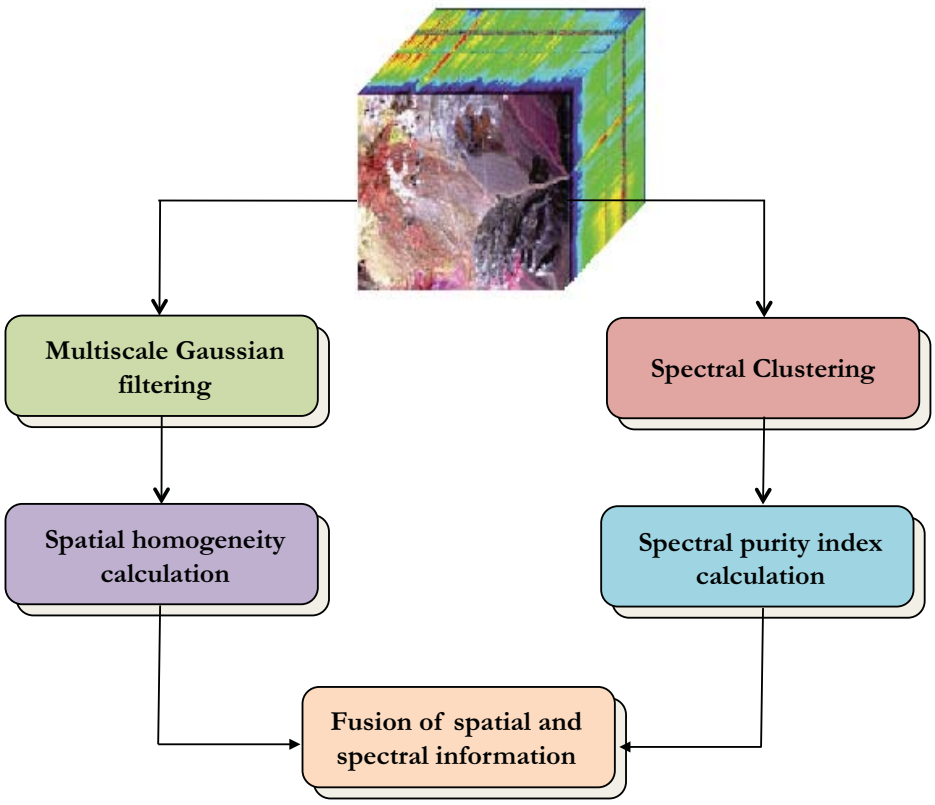


Figure 1: Block diagram illustrating the spatial-spectral preprocessing (SSPP) method.

 ISBN: 978-84-608-6082-2

results in a filtered version $\mathbf{Y}_F$ of the original hyperspectral image. Let us denote by $\mathbf{y}(i,j) = [y_1(i,j), y_2(i,j), \cdots, y_B(i,j)]$ the B-dimensional pixel vector at spatial coordinates (i,j) of the hyperspectral image $\mathbf{Y}$, which can now be defined as a set $\mathbf{Y} = \{\mathbf{y}(i,j)\}_{i \in 1,\cdots,r}$. Eq. (1) shows the pixel-level operation that we perform for each k-th spectral band of the hyperspectral image, with $1 \leq k \leq B$:

$$F_k[\mathbf{y}(i,j)] = \sum_{i'=1}^{r} \sum_{j'=1}^{c} G(i-i', j-j') \cdot y_k(i',j'), \text{ with } G(i',j') = \frac{1}{2\pi\sigma^2} e^{-\frac{i'^2+j'^2}{2\sigma^2}}. \quad (1)$$

Spatial homogeneity calculation. This step takes as input the filtered hyperspectral image obtained in the previous step and produces a spatial homogeneity index for each pixel in the original image $\mathbf{Y}$. To perform this step, we first calculate the root mean square error (RMSE) [6] between the original hyperspectral image and the filtered image. Eq. (2) indicates the operation to calculate the RMSE between the pixel $\mathbf{y}(i,j)$ in the original image and the pixel at the same spatial coordinates, $\mathbf{y}_F(i,j)$, in the filtered image:

$$\text{RMSE}[\mathbf{y}(i,j), \mathbf{y}_F(i,j)] = \left(\frac{1}{B} \sum_{k=1}^{B} (y_k(i,j) - y_{F_k}(i,j))^2 \right)^{\frac{1}{2}}. \quad (2)$$

The lower the RMSE score, the higher the similarity between the pixel in the original image and its neighbors. Quite opposite, the higher the RMSE, the lower the similarity of the pixel in the original image with regards to its neighbours. As a result, the RMSE in Eq. (2) can be used as a spatial homogeneity index for each pixel $\mathbf{y}(i,j)$ in the hyperspectral image $\mathbf{Y}$.

In the *spectral purity index calculation* step, we first use principal component analysis (PCA) [7] to reduce the dimensionality of the hyperspectral image, retaining the first p principal components (PCs) containing most of the variance in the data. Then, we use the first PCs as the skewers for which we identify the pixels with maxima and minima projection values, following a procedure similar to the one adopted by the pixel purity index (PPI) algorithm in [8]. The pixels with maxima and minima projection values are assigned a weight of 1. The weight of the mean value between the maxima and minima projection value is 0. A threshold value is also applied so that the weights lower than this threshold are assigned the value 0. Finally the spectral purity is calculated as the sum of all the weights over the first p PCs.

In the *spectral clustering*, we perform a spectral-based unsupervised clustering of the original hyperspectral image. This step, which is applied separately from the previous steps, uses the K-Means algorithm [9] in order to identify p clusters in the hyperspectral image.

Fusion of spatial and spectral information is a step that takes as input the spatial homogeneity index calculated in the second step and the clusters calculated in fourth step,

and returns a subset of candidate pixels in the original hyperspectral image which will be used for endmember identification purposes. For each cluster, a subset of spatially homogeneous and spectrally pure pixels is selected. To do so, pixels in each cluster are ranked according to increasing values of their spatial homogeneity and spectral purity.

Finally, an endmember extraction algorithm can be applied to the pixels retained after the procedure above. The outcome of the process is a set of p endmembers and their corresponding fractional abundance maps (one per endmember).

3 Parallel Implementation

The parallel implementation of SSPP has been developed using OpenMP which is an API used to explicity address multithreaded, shared-memory parallelism. In OpenMP the users specify the region in the code that are suitable for parallel implementation using pragmas and clauses supported in the gnu or Intel compilers. In the following, we briefly summarize the main techniques used in the multi-core implementation of the considered algorithm:

1. *Multi-scale Gaussian filtering.* We have developed an optimization where the most consuming part using the convolution of two 1-dimensional filters is the central part. This optimization consists on declare a *#pragma omp parallel for schedule (static, 32)* where the loop is divided into 32 equal-size chunks and also, an unrolling is developed to vectorize this central part to calculate the gaussian filtering. We have empirically tested that, if the size of the secene is largest, this value should be increased.

2. *Spatial homogeneity calculation.* The RMSE is required to calculate the spatial homogeneity index for each pixel in the data set. A reduction process for each pixel is computed where *#pragma omp simd* is applied to vectorize the reduction whose main loop is based on each spectral band. Finally, for each pixel we have used a *#pragma omp parallel for* to divide the loop iterations between the spawned threads and compute the square for each pixel in the scene.

3. *Spectral purity index calculation.* The PCA operation is required to calculate the spectral purity index. First of all, we need to calculate the normalized image obtained by substracting the average of all pixels in the scene to each pixel in the original image. For this purpose, the reduction process is performed using the *#pragma omp parallel for private (mean)*, where *mean* is the average of all pixels in the scene and later. For this calculation, *#pragma omp simd* is used to vectorize the loop. The same strategy is applied to substract the *mean* value to each pixel in the data set. After that, *mkl and lapack* are used to compute matrix multiplications to obatin the reduced image.

4. *Spectral clustering and fusion of spatial and spectral information.* For both steps, we have not applied any parallel technique to accelerate the process because the

processing time is lower.

4 Experimental validation

The experiments are carried out using a collection of 24 synthetic hyperspectral images simulated with different sizes (10000 to 200000 pixels) and number of endmembers (10 to 30). The signatures are obtained from the USGS library and the scenes are generated using the procedure described in [10] to simulate natural spatial patterns. These images comprise 224 narrow spectral bands between 0.4 to 2.5 μm. On the other hand, we have used the well-known AVIRIS Cuprite scene, collected by the Airborne Visible Infra-Red Imaging Specrometer (AVIRIS) in the summer of 1997 and available online in reflectance units after atmospheric correction. The portion comprises a relatively large area (350 lines by 350 samples and 20-m pixels) and 224 spectral bands between 0.4 to 2.5 μm and a total size of around 46 MB. Bands 1-3, 105-115, and 150-170 were removed prior to the analysis due to water absorption and low SNR in those bands prior to the analysis.

In order to evalute the performance, the proposed multi-core implementation has been tested on the following platform: 2×Intel Xeon processor E5-2670 with 8 cores each, at 2.60 GHz and 32 GB of DDR3 RAM memory. The Figs. 2 and 3 show the processing times considering different number of endmembers and execution threads. As can be seen, increasing the number of threads allows a better parallel performance and speedup respect to our single-threaded optimized implementation. On the other hand, Fig. 4 shows the execution times obtained for the AVIRIS Cuprite scene.

For illustrative purposes, Table 1 shows the timming results and speedups for each data set used in the experiments. As shown in Table 1, our data sets could be processed with significant speedup factor using the Intel Xeon, up to 3 times.

Table 1: Mean execution times (in seconds) and speedups (in the parentheses) for the best multi-core setting using real and synthetic scenes with different sizes and endmembers.

Image	**10 endmembers**	**20 endmembers**	**30 endmembers**
100×100 - 8 threads	0.0697 (1.68)	0.0697 (1.68)	0.0704 (1.72)
200×100 - 8 threads	0.1136 (1.81)	0.1101 (1.81)	0.1120 (1.82)
300×100 - 8 threads	0.1576 (1.85)	0.1490 (1.93)	0.1483 (1.94)
400×100 - 8 threads	0.1805 (2.02)	0.1830 (1.89)	0.1801 (2.06)
100×500 - 8 threads	0.2112 (2.13)	0.2027 (2.12)	0.2034 (2.07)
200×100 - 8 threads	0.3649 (2.25)	0.3582 (2.22)	0.3591 (2.20)
300×100 - 8 threads	0.5104 (2.26)	0.5028 (2.18)	0.5115 (2.22))
400×100 - 8 threads	0.6494 (2.39)	0.6568 (2.38)	0.6729 (2.39)
AVIRIS Cuprite - 14 threads	0.3579 (2.72) - 19 endmembers		

 ISBN: 978-84-608-6082-2

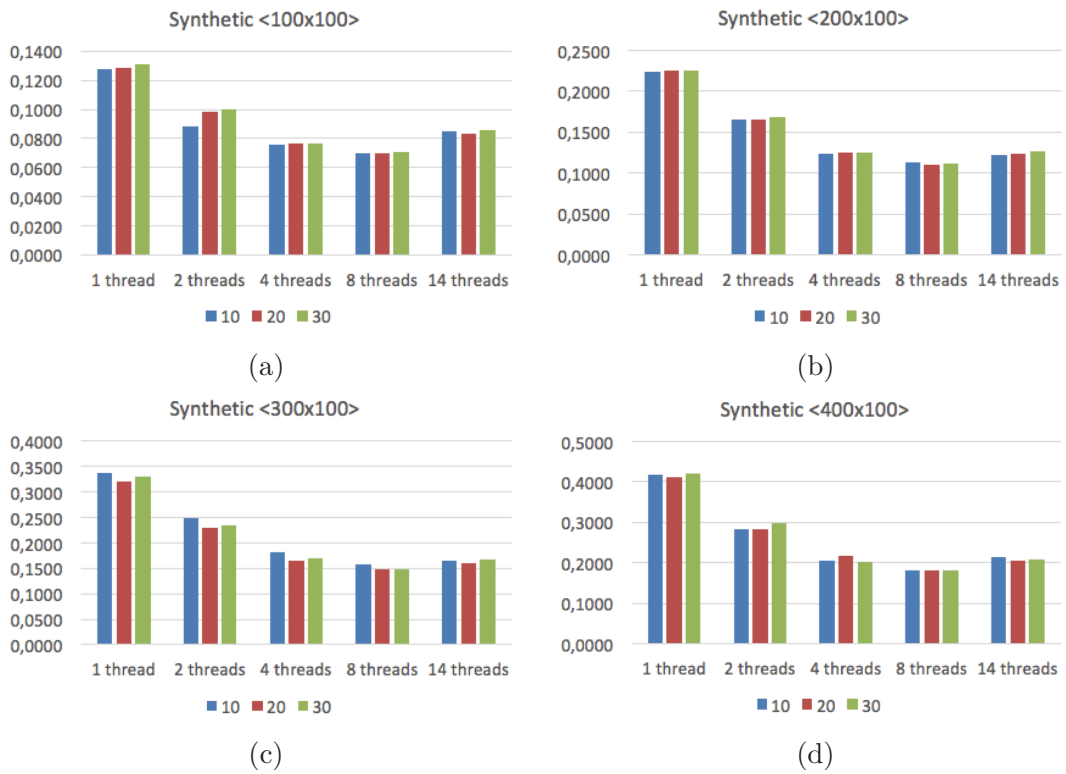


Figure 2: Execution times (seconds) for the multi-core version of the SSPP algorithm, considering different number of endmembers and sizes: (a) 100×100 pixels. (b) 200×100 pixels. (c) 300×100 pixels. (d) 400×100 pixels.

 ISBN: 978-84-608-6082-2

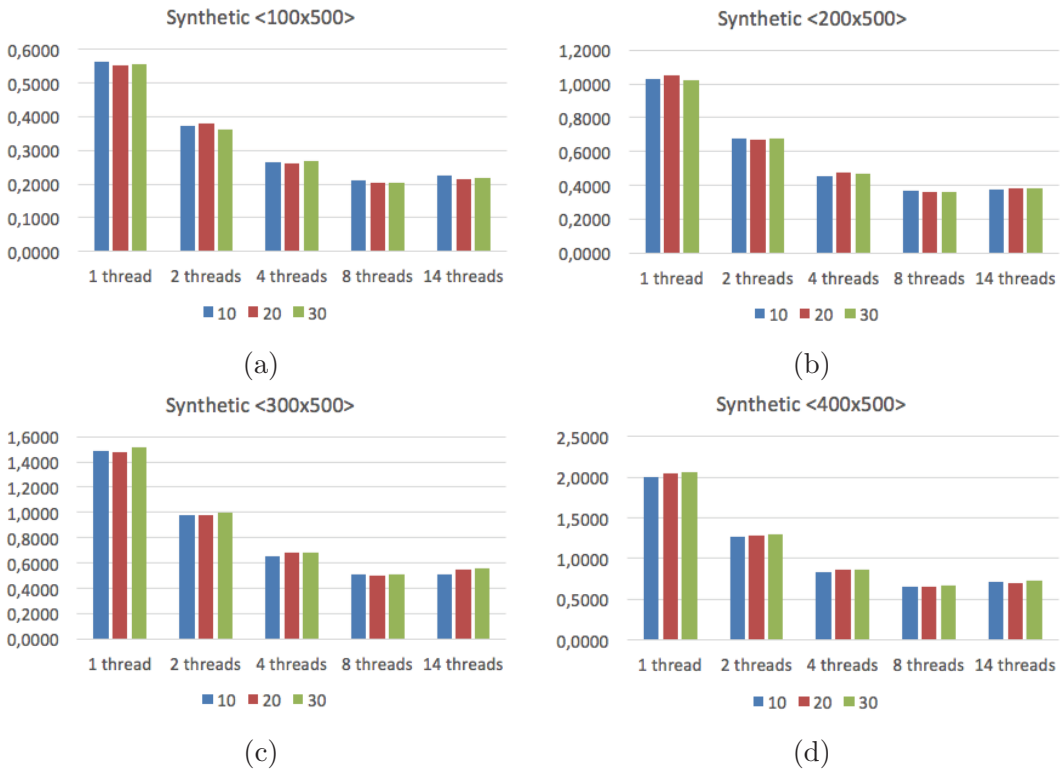


Figure 3: Execution times (seconds) for the multi-core version of the SSPP algorithm, considering different number of endmembers and sizes: (a) 100×500 pixels. (b) 200×500 pixels. (c) 300×500 pixels. (d) 400×500 pixels.

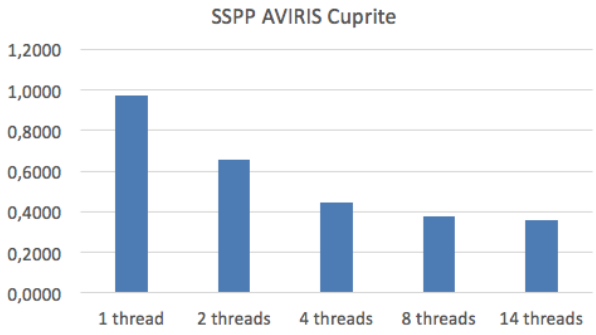


Figure 4: Execution times (seconds) for the multi-core version of the SSPP algorithm, considering the AVIRIS Cuprite scene.

5 Conclusions

A multi-core implementation of the SSPP algorithm for spectral unmixing has been proposed. The optimized version exploits different strategies using OpenMP: vectorization and shared memory between threads. The obtained results indicate that significant performance could be obtained using an Intel Xeon processor E5-2670 platform. Further experimentation with additional real scenes and a comparison with another programming languages are desirable in future research developments.

Acknowledgements

This work has been supported by Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005) and by the Formación Posdoctoral programme (FPDI-2013-16280). This work was partially supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain. Funding from the Spanish Ministry of Economy and Competitiveness (MINECO) through the research contracts TIN2012-32180 and TIN2015-65277-R are also gratefully acknowledged.

References

[1] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[2] J. Plaza, E. M. T. Hendrix, I. Garcia, G. Martin, and A. Plaza, "On endmember identification in hyperspectral images without pure pixels: A comparison of algorithms," *Journal of Mathematical Imaging and Vision*, vol. 42, no. 2-3, pp. 163–175, 2012.

[3] M. Zortea and A. Plaza, "Spatial preprocessing for endmember extraction," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 8, pp. 2679–2693, 2009.

[4] G. Martin and A. Plaza, "Region-based spatial preprocessing for endmember extraction and spectral unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 4, pp. 745–749, 2011.

[5] G. Martin and A. Plaza., "Spatial-spectral preprocessing prior to endmember identification and unmixing of remotely sensed hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 5, no. 2, pp. 380–395, 2012.

[6] N. Keshava and J. F. Mustard, "Spectral unmixing," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 44–57, 2002.

[7] J. A. R. y X. Jia, "Remote Sensing Digital Image Analysis," in *Springer- Verlag*, 1999.

[8] J. W. Boardman, F. A. Kruse, and R. O. Green, "Mapping Target Signatures Via Partial Unmixing of Aviris Data," *Proc. JPL Airborne Earth Sci. Workshop*, pp. 23–26, 1995.

[9] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[10] G. S. Miller, "The definition and rendering of terrain maps," in *ACM SIGGRAPH Computer Graphics*, vol. 20, no. 4. ACM, 1986, pp. 39–48.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Widely Linear Quaternion Signal Filter from One-Step Delayed Observations

J. D. Jiménez-López[1], R. M. Fernández-Alcalá[1], J. C. Ruiz-Molina[1] and J. Navarro-Moreno[1]

[1] *Department of Statistics and Operations Research, University of Jaén, Spain*

emails: jdomingo@ujaen.es, rmfernan@ujaen.es, jcruiz@ujaen.es, jnavarro@ujaen.es

Abstract

This paper is concerned with the optimal quaternion filtering problem for linear discrete-time stochastic systems with one-step delayed observations. A recursive algorithm for the widely linear filter is proposed, derived from the augmented state-space model formed by the quaternion signal and observation together with their involutions.

Key words: delayed observations, filter, quaternion signals, widely linear processing

1 Introduction

Recently, in many research papers on signal estimation, the problem is approached by using quaternion domain, thus giving rise to that the optimal linear processing is full-widely linear, from now on called widely linear (WL). Its main characteristic is that, for the estimation, it is used not only the quaternion vector but also its three involutions, allows us to obtain considerable improvement over the conventional optimal processing. Several estimation algorithms has been newly developed under a WL processing. We will mention the research papers [**?**]-[**?**], in which adaptive filters are obtained, or [**?**], in which a Kalman filtering algorithm is proposed by using quaternion algebra, or [**?**]-[**?**], in which the estimation problem is approached under the semi-widely linear (SWL) processing, consisting of the use of the quaternion $\mathbb{C}^{\eta}$-properness. Notice that none of the above WL or SWL solutions incorporate delayed observations.

In the last years, since its application to real situations, there has also been a proliferation of research papers approaching signal estimation problems from observations which can

 ISBN: 978-84-608-6082-2

be updated or delayed at each instant of time (see, for example, [**?**],[**?**] and the references therein). All the references before consider real signal and delayed observations, but in this paper, we propose a recursive filtering algorithm to obtain the WL signal estimator from observations updated or one-step delayed at each instant of time in quaternion domain.

2 Problem statement

All the random variables in this paper are assumed of zero-mean. Let $x_n \in \mathbb{H}$ be a quaternion state, given by $x_n = x_{n,r} + \eta x_{n,\eta} + \eta' x_{n,\eta'} + \eta'' x_{n,\eta''}$, where $x_{n,r}$, $x_{n,i}$, $x_{n,j}$ and $x_{n,k}$ are real random signals and $\{1, \eta, \eta', \eta''\}$ is a quaternion orthogonal basis. It verifies this equation:

$$x_n = f_n x_{n-1} + g_n x_{n-1}^{\eta} + h_n x_{n-1}^{\eta'} + e_n x_{n-1}^{\eta''} + w_{n-1}, \quad n \geq 1 \tag{1}$$

where f_n, g_n, h_n, e_n are deterministic and w_n is a quaternion white noise. Similarly, assume that the real quaternion observations $\tilde{z}_n \in \mathbb{H}$ satisfy the following equation:

$$\tilde{z}_n = x_n + v_n, \quad n \geq 1 \tag{2}$$

where v_n a quaternion white noise. Let us also consider that, at each instant of time, every quaternion component of the available observation can coincide with the corresponding component of the real observation at this time (*updated*) or be *delayed* one sampling period. To model this situation, let us consider the following observation equation

$$z_n = (1 - \gamma_n) \star \tilde{z}_n + \gamma_n \star \tilde{z}_{n-1}, \quad n \geq 2 \tag{3}$$

where $\gamma_n \in \mathbb{H}$ with components $\gamma_{n,l}$, for $l = r, \eta, \eta', \eta''$ which are independent Bernoulli random variables with parameters $p_{n,l}$, and $\star$ denotes the product $\star$ between two quaternions $x_n, y_m \in \mathbb{H}$, defined as $x_n \star y_m := x_{n,r} y_{m,r} + \eta x_{n,\eta} y_{m,\eta} + \eta' x_{n,\eta'} y_{m,\eta'} + \eta'' x_{n,\eta''} y_{m,\eta''}$. Observation equation (**??**) ensures that the first available observation is updated ($z_1 = \tilde{z}_1$), and from the second one, their quaternion components can be updated or delayed one time.

From this model, our aim is to obtain recursive algorithms for calculating the WL filter of the signal x_n from the observations $\{z_1, \ldots, z_n\}$, denoted by $\hat{x}_{n/n}^{\mathrm{WL}}$. For this purpose, we first derive the augmented WL system containing the state and the observations, together with their involutions. Taking into account the properties of the quaternion involutions, from (**??**) the following augmented state equation is obtained:

$$\bar{\mathbf{x}}_n = \mathbf{F}_n \bar{\mathbf{x}}_{n-1} + \bar{\mathbf{w}}_{n-1} \tag{4}$$

where $\bar{\mathbf{x}}_n = \left[x_n, x_n^{\eta}, x_n^{\eta'}, x_n^{\eta''}\right]^T$, $\bar{\mathbf{w}}_n = \left[w_n, w_n^{\eta}, w_n^{\eta'}, w_n^{\eta''}\right]^T$ (T denotes the transpose vector) and

$$\mathbf{F}_n = \begin{bmatrix} f_n & g_n & h_n & e_n \\ g_n^{\eta} & f_n^{\eta} & e_n^{\eta} & h_n^{\eta} \\ h_n^{\eta'} & e_n^{\eta'} & f_n^{\eta'} & g_n^{\eta'} \\ e_n^{\eta''} & h_n^{\eta''} & g_n^{\eta''} & f_n^{\eta''} \end{bmatrix}$$

Between the augmented vector of x_n and the real vector $\mathbf{x}_n^r = [x_{n,r}, x_{n,\eta}, x_{n,\eta'}, x_{n,\eta''}]^T$ can be established the following relation $\bar{\mathbf{x}}_n = \boldsymbol{\mathcal{A}}\mathbf{x}_n^r$, where

$$\boldsymbol{\mathcal{A}} = \begin{bmatrix} 1 & \eta & \eta' & \eta'' \\ 1 & \eta & -\eta' & -\eta'' \\ 1 & -\eta & \eta' & -\eta'' \\ 1 & -\eta & -\eta' & \eta'' \end{bmatrix}$$

with $\boldsymbol{\mathcal{A}}^H\boldsymbol{\mathcal{A}} = 4\mathbf{I}_4$ (H denotes the Hermitian vector and $\mathbf{I}_n$ the $n \times n$ identity matrix).

Additionally, we denote the covariance matrix of the initial state vector $\bar{\mathbf{x}}_0$ by $\mathbf{P}_0$, and we assume that $\bar{\mathbf{w}}_n$ is a quaternion white noise vector with $E\left[\bar{\mathbf{w}}_n\bar{\mathbf{w}}_n^H\right] = \mathbf{Q}_n$.

Similarly, from (**??**), (**??**), the augmented observation equation can be written as

$$\begin{array}{c} \bar{\mathbf{z}}_n = (\mathbf{I}_4 - \boldsymbol{\Gamma}_n)\,\bar{\tilde{\mathbf{z}}}_n + \boldsymbol{\Gamma}_n\bar{\tilde{\mathbf{z}}}_{n-1},\ n \geq 2; \quad \bar{\mathbf{z}}_1 = \bar{\tilde{\mathbf{z}}}_1 \\ \bar{\tilde{\mathbf{z}}}_n = \bar{\mathbf{z}}_n + \bar{\mathbf{v}}_n \end{array} \tag{5}$$

where $\boldsymbol{\Gamma}_n = \frac{1}{4}\boldsymbol{\mathcal{A}}\mathrm{diag}(\boldsymbol{\gamma}_n^r)\boldsymbol{\mathcal{A}}^H$, with $\mathrm{diag}(\cdot)$ denoting a diagonal matrix with the elements specified on the main diagonal.

Moreover, we assume that $\bar{\mathbf{v}}_n$ is a quaternion white noise vector with $E\left[\bar{\mathbf{v}}_n\bar{\mathbf{v}}_n^H\right] = \mathbf{R}_n$ and also that $\bar{\mathbf{x}}_n, \bar{\mathbf{w}}_n, \bar{\mathbf{v}}_n$ and $\boldsymbol{\Gamma}_n$, are mutually independent.

3 WL filtering algorithm

For the system (**??**)-(**??**), the WL filter is obtained as $\hat{x}_{n/n}^{\mathrm{WL}} = [1,0,0,0]\hat{\bar{\mathbf{x}}}_{n/n}$, where $\hat{\bar{\mathbf{x}}}_{n/n}$ is calculated as

$$\begin{array}{l} \hat{\bar{\mathbf{x}}}_{n/n} = \mathbf{F}_n\hat{\bar{\mathbf{x}}}_{n-1/n-1} + \mathbf{S}_{n,n}^{\bar{\mathbf{x}}}\boldsymbol{\Omega}_n^{-1}\boldsymbol{\nu}_n, \quad n \geq 2 \\ \hat{\bar{\mathbf{x}}}_{1/1} = \left(\mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^H + \mathbf{Q}_0\right)\boldsymbol{\Omega}_1^{-1}\bar{\mathbf{z}}_1 \end{array}$$

$\boldsymbol{\nu}_n$ vectors are given by

$$\begin{array}{rl} \boldsymbol{\nu}_n = & \bar{\mathbf{z}}_n - \left((\mathbf{I}_4 - \boldsymbol{\Pi}_n)\,\mathbf{F}_n + \boldsymbol{\Pi}_n\right)\hat{\bar{\mathbf{x}}}_{n-1/n-1} - \boldsymbol{\Pi}_n\left(\mathbf{I}_4 - \boldsymbol{\Pi}_{n-1}\right)\mathbf{R}_{n-1}\boldsymbol{\Omega}_{n-1}^{-1}\boldsymbol{\nu}_{n-1}, \quad n \geq 2 \\ \boldsymbol{\nu}_1 = & \bar{\mathbf{z}}_1 \end{array}$$

with $\boldsymbol{\Pi}_n = \dfrac{1}{4}\boldsymbol{\mathcal{A}}\mathrm{diag}(E[\boldsymbol{\gamma}_n^r])\boldsymbol{\mathcal{A}}^H$, for $n \geq 1$, and $\boldsymbol{\Omega}_n$ is of the form

$$\begin{array}{rl} \boldsymbol{\Omega}_n = & \boldsymbol{\mathcal{A}}\mathrm{diag}\left(\mathrm{var}\left(\boldsymbol{\gamma}_n^r\right) \circ \left(E\left[\left(\mathbf{x}_n^r - \mathbf{x}_{n-1}^r\right)\left(\mathbf{x}_n^r - \mathbf{x}_{n-1}^r\right)^H\right] + E\left[\mathbf{v}_n^r\mathbf{v}_n^{rH}\right]\right)\right)\boldsymbol{\mathcal{A}}^H \\ & + \left((\mathbf{I}_4 - \boldsymbol{\Pi}_n)\,\mathbf{F}_n + \boldsymbol{\Pi}_n\right)\mathbf{P}_{n-1,n-1/n-1}^{\bar{\mathbf{x}}\bar{\mathbf{x}}}\left((\mathbf{I}_4 - \boldsymbol{\Pi}_n)\,\mathbf{F}_n + \boldsymbol{\Pi}_n\right)^H \\ & + \left(\mathbf{I}_4 - \boldsymbol{\Pi}_n\right)\left(\mathbf{Q}_{n-1} + \mathbf{R}_n\right)\left(\mathbf{I}_4 - \boldsymbol{\Pi}_n\right) \\ & + \boldsymbol{\Pi}_n\left(\mathbf{R}_{n-1} - \mathbf{R}_{n-1}\boldsymbol{\Omega}_{n-1}^{-1}\left(\mathbf{I}_4 - \boldsymbol{\Pi}_{n-1}\right)\mathbf{R}_{n-1}\right)\boldsymbol{\Pi}_n, \quad n \geq 2 \\ \boldsymbol{\Omega}_1 = & \mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^H + \mathbf{Q}_0 + \mathbf{R}_1 \end{array}$$

being $\circ$ the Hadamard product.

Matrix $\mathbf{S}_{n,n}^{\bar{\mathbf{x}}}$ is calculated as follows

$$\begin{array}{rl}\mathbf{S}_{n,n}^{\bar{\mathbf{x}}} = & \mathbf{F}_n\mathbf{P}_{n-1,n-1/n-1}^{\bar{\mathbf{x}}\bar{\mathbf{x}}}\left(\left(\mathbf{I}_4-\mathbf{\Pi}_n\right)\mathbf{F}_n+\mathbf{\Pi}_n\right)^H+\mathbf{Q}_{n-1}\left(\mathbf{I}_4-\mathbf{\Pi}_n\right), \quad n\geq 2\\ \mathbf{S}_{1,1}^{\bar{\mathbf{x}}} = & \mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^H+\mathbf{Q}_0\end{array}$$

The WL filtering error is computed by $p_{n,n/n}^{\mathrm{WL}}=[1,0,0,0]\mathbf{P}_{n,n/n}^{\bar{\mathbf{x}}\bar{\mathbf{x}}}[1,0,0,0]^T$ where

$$\begin{array}{rl}\mathbf{P}_{n,n/n}^{\bar{\mathbf{x}}\bar{\mathbf{x}}} = & \mathbf{F}_n\mathbf{P}_{n-1,n-1/n-1}^{\bar{\mathbf{x}}\bar{\mathbf{x}}}\mathbf{F}_n^H+\mathbf{Q}_{n-1}-\mathbf{S}_{n,n}^{\bar{\mathbf{x}}}\mathbf{\Omega}_n^{-1}\mathbf{S}_{n-1,n}^{\bar{\mathbf{x}}H}, \quad n\geq 2\\ \mathbf{P}_{1,1/1}^{\bar{\mathbf{x}}\bar{\mathbf{x}}} = & \left(\mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^H+\mathbf{Q}_0\right)\left[\mathbf{I}_4-\mathbf{\Omega}_1^{-1}\left(\mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^H+\mathbf{Q}_0\right)^H\right]\end{array}$$

References

[1] C. Ch. Took and D. P. Mandic. *A quaternion widely linear adaptive filter*, IEEE Trans. Signal Process. **58(8)** (2010) 4427–4431.

[2] B. Che Ujang, C. Ch. Took and D. P. Mandic. *Quaternion-valued nonlinear adaptive filtering*, IEEE Trans. Neural Netw. **22(8)** (2011) 1193–1206.

[3] C. Jahanchahi, C. Ch. Took and D. P. Mandic. *A class of quaternion-valued affine projection algorithms*, Signal Process. **93(7)** (2013) 1712–1723.

[4] C. Jahanchahi and D. P. Mandic. *A class of quaternion Kalman filters*, IEEE Trans. Neural Netw. Learn. Syst. **25(3)** (2014) 533–544.

[5] J. Navarro-Moreno, R. M. Fernández-Alcalá and J. C. Ruiz-Molina. *A quaternion widely linear model for nonlinear Gaussian estimation*, IEEE Trans. Signal Process. **62(24)** (2014) 6414–6424.

[6] J. Navarro-Moreno, R. M. Fernández-Alcalá and J. C. Ruiz-Molina. *Semi-widely simulation and estimation of continuous-time $\mathbb{C}^\eta$-proper quaternion random signals*, IEEE Trans. Signal Process. **63(18)** (2015) 4999–5012.

[7] J. Navarro-Moreno and J. C. Ruiz-Molina. *Semi-widely linear estimation of $\mathbb{C}^\eta$-proper quaternion random signal vectors under Gaussian and stationary conditions*, Signal Process. **119** (2016) 56–66.

[8] D. Chen, L. Xu and J. Du. *Optimal filtering for systems with finite-step autocorrelated process noises, random one-step sensor delay and missing measurements*, Commun. Nonlinear Sci. **32** (2016) 211–224.

[9] S. Wang, H. Fang and X. Tian. *Recursive estimation for nonlinear stochastic systems with multi-step transmission delays, multiple packet dropouts and correlated noises*, Signal Process. **115** (2015) 164–175.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An extension of the Muth distribution

P. Jodrá[1], H.W. Gómez[2], M.D. Jiménez-Gamero[3] and M.V. Alba-Fernández[4]

[1] *Departamento de Métodos Estadísticos, Universidad de Zaragoza*

[2] *Departamento de Matemáticas, Universidad de Antofagasta*

[3] *Departamento de Estadística e Investigación Operativa, Universidad de Sevilla*

[4] *Departamento de Estadística e Investigación Operativa, Universidad de Jaén*

emails: pjodra@unizar.es, hector.gomez@uantof.cl, dolores@us.es, mvalba@ujaen.es

Abstract

Muth introduced a probability distribution with application in reliability theory. This paper proposes an extension of this distribution and studies some statistical properties of the new model, such as the computation of the moments, computer generation of pseudo-random data and the behaviour of the failure rate function. The estimation of parameters is carried out by the method of maximum likelihood. The practical usefulness of the model is illustrated by a real data set, showing that it may provide a better fit than other previously considered probability distributions.

Key words: Muth distribution, generalized integro-exponential function, Lambert W function, point estimation, data analysis
MSC 2000: 60E05, 62F10

1 Introduction

A random variable Y is said to have a Muth distribution if its probability density function (pdf) is given by

$$f_Y(y;\alpha) = (e^{\alpha y} - \alpha)\, \exp\left\{\alpha y - \frac{1}{\alpha}\,(e^{\alpha y} - 1)\right\}, \qquad y > 0,$$

where $\alpha \in (0,1]$ is a shape parameter. The cumulative distribution function (cdf) of Y is the following

$$F_Y(y;\alpha) = 1 - \exp\left\{\alpha y - \frac{1}{\alpha}\left(e^{\alpha y} - 1\right)\right\}, \qquad y > 0. \tag{1}$$

As a natural extension, Jodrá et al. [4] considered the scaled Muth distribution, which is defined as βY, with $\beta > 0$, and they showed the usefulness of the scaled Muth law for modelling rainfall data.

Here it is introduced a new probability distribution from the Muth distribution, which includes it as a particular case. More specifically, it is defined the power Muth (PM) distribution by means of the transformation $X = \beta Y^{1/\gamma}$, where Y is the Muth law with parameter $\alpha \in (0,1]$, $\beta > 0$ and $\gamma > 0$. The introduction of the new parameter γ leads to a richer class of probability distributions for non-negative random variables. with a wide range of values for the asymmetry and kurtosis coefficients, increasing generalized failure rate as well as increasing or bathtub shape failure rate.

Using ordinary results related to the transformation of variables, it is easy to see that that the pdf and cdf of X are given, respectively, by

$$f(x;\alpha,\beta,\gamma) = \frac{\gamma}{\beta^\gamma}\, x^{\gamma-1}\left(e^{\alpha\left(\frac{x}{\beta}\right)^\gamma} - \alpha\right)\exp\left\{\alpha\left(\frac{x}{\beta}\right)^\gamma - \frac{1}{\alpha}\left(e^{\alpha\left(\frac{x}{\beta}\right)^\gamma} - 1\right)\right\}, \qquad x > 0,$$

and

$$F(x;\alpha,\beta,\gamma) = 1 - \exp\left\{\alpha\left(\frac{x}{\beta}\right)^\gamma - \frac{1}{\alpha}\left(e^{\alpha\left(\frac{x}{\beta}\right)^\gamma} - 1\right)\right\}, \qquad x > 0, \tag{2}$$

where α and γ are shape parameters and β is a scale parameter. For the sake of brevity, the power Muth distribution will be denoted by $\mathrm{PM}(\alpha,\beta,\gamma)$.

2 Statistical properties

2.1 Moments

The moments of the PM distribution can be expressed in terms of the generalized integro-exponential function, which is defined by the following integral representation (see Milgram [5])

$$E_s^m(z) = \frac{1}{\Gamma(m+1)}\int_1^\infty (\log u)^m e^{-zu}u^{-s}du, \qquad z \in (-\infty,\infty), \tag{3}$$

where $s \in (-\infty,\infty)$ and $m > -1$, Γ stands for the Gamma function and log for the natural logarithm. With the above notation, we establish the following.

Proposition 1 *Let X be a $\mathrm{PM}(\alpha,\beta,\gamma)$ distribution. Then,*

$$E[X^k] = \frac{\beta^k e^{1/\alpha}}{\alpha^{k/\gamma}}\,\Gamma\left(\frac{k}{\gamma}+1\right)E_0^{\frac{k}{\gamma}-1}\left(\frac{1}{\alpha}\right), \qquad k = 1,2,\ldots. \tag{4}$$

 ISBN: 978-84-608-6082-2

2.2 Quantile function

The PM distribution inherits the variate generation property from the Muth distribution, that is, its quantile function can also be given in closed form. Specifically, it can be expressed explicitly in terms of the Lambert W function (see Corless et al. [3] for a review of the theory and applications of W). In this regard, it is interesting to note that the Lambert W function is implemented in computer algebra systems and programming languages such as R [6]. Therefore, pseudo-random data from the PM model can be computer-generated in a straightforward manner by virtue of the following result.

Proposition 2 *The quantile function of the* $\mathrm{PM}(\alpha,\beta,\gamma)$ *distribution,* $F^{-1}(u;\alpha,\beta,\gamma)$*, is given by*

$$F^{-1}(u;\alpha,\beta,\gamma)=\beta\left(\frac{1}{\alpha}\,\log\left\{-\alpha W_{-1}\left(\frac{u-1}{\alpha e^{1/\alpha}}\right)\right\}\right)^{1/\gamma},\qquad 0<u<1, \tag{5}$$

where W_{-1} *denotes the negative branch of the Lambert* W *function.*

2.3 Failure rate function

The failure (or hazard) rate function of the $\mathrm{PM}(\alpha,\beta,\gamma)$ distribution is given by

$$h(x;\alpha,\beta,\gamma)=\frac{f(x;\alpha,\beta,\gamma)}{1-F(x;\alpha,\beta,\gamma)}=\frac{\gamma}{\beta^{\gamma}}\left(e^{\alpha\left(\frac{x}{\beta}\right)^{\gamma}}-\alpha\right)x^{\gamma-1},\qquad x>0.$$

Proposition 3 (a) *If* $\gamma\geq 1$ *then* $h(x;\alpha,\beta,\gamma)$ *is increasing in* x *for any* $\alpha\in(0,1]$ *and* $\beta>0$.
(b) *If* $0<\gamma<1$ *then there exists an* $x_0=x_0(\alpha,\beta,\gamma)>0$ *so that* $h(x;\alpha,\beta,\gamma)$ *is (strictly) decreasing when* $x<x_0$ *and (strictly) increasing when* $x>x_0$.

Therefore, the $\mathrm{PM}(\alpha,\beta,\gamma)$ distribution has either increasing failure rate when $\gamma\geq 1$ or bathtub-shaped failure rate when $0<\gamma<1$.

3 Parameter estimation

Let $X_1,\ldots,X_n$ be a random sample of size n from the $\mathrm{PM}(\alpha,\beta,\gamma)$ distribution with unknown parameters α, β and γ. Denote by $x_1, x_2, \ldots, x_n$ the observed values of the sample. From the likelihood function, $L(\alpha,\beta,\gamma)=\prod_{i=1}^{n} f(x_i;\alpha,\beta,\gamma)$, the log-likelihood function

can be written as follows

$$\begin{aligned}\log L(\alpha,\beta,\gamma) &= n(\log\gamma-\gamma\log\beta)+(\gamma-1)\sum_{i=1}^{n}\log x_i+\sum_{i=1}^{n}\log\left\{e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-\alpha\right\}\\ &\quad+\frac{\alpha}{\beta^{\gamma}}\sum_{i=1}^{n}x_i^{\gamma}-\frac{1}{\alpha}\sum_{i=1}^{n}\left(e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-1\right).\end{aligned}$$

The maximum likelihood (ML) estimates of α,β,γ are the values $\hat{\alpha},\hat{\beta},\hat{\gamma}$ that maximize the log-likelihood function. The system of partial derivatives of $\log L(\alpha,\beta,\gamma)$ is the following

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log L &= \sum_{i=1}^{n}\frac{(\frac{x_i}{\beta})^{\gamma}e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-1}{e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-\alpha}-\frac{1}{\alpha\beta^{\gamma}}\sum_{i=1}^{n}x_i^{\gamma}e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}+\frac{1}{\alpha^2}\sum_{i=1}^{n}e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}\\ &\quad+\frac{1}{\beta^{\gamma}}\sum_{i=1}^{n}x_i^{\gamma}-\frac{n}{\alpha^2}=0,\\ \frac{\partial}{\partial\beta}\log L &= -\frac{\alpha\gamma}{\beta^{\gamma+1}}\sum_{i=1}^{n}\frac{x_i^{\gamma}e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}}{e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-\alpha}+\frac{\gamma}{\beta^{\gamma+1}}\sum_{i=1}^{n}x_i^{\gamma}e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-\frac{\alpha\gamma}{\beta^{\gamma+1}}\sum_{i=1}^{n}x_i^{\gamma}-\frac{n\gamma}{\beta}=0,\\ \frac{\partial}{\partial\gamma}\log L &= \frac{\alpha}{\beta^{\gamma}}\sum_{i=1}^{n}\frac{x_i^{\gamma}\log\left(\frac{x_i}{\beta}\right)e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}}{e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}-\alpha}-\frac{1}{\beta^{\gamma}}\sum_{i=1}^{n}x_i^{\gamma}\log\left(\frac{x_i}{\beta}\right)e^{\alpha\left(\frac{x_i}{\beta}\right)^{\gamma}}\\ &\quad+\frac{\alpha}{\beta^{\gamma}}\sum_{i=1}^{n}x_i^{\gamma}\log\left(\frac{x_i}{\beta}\right)+\sum_{i=1}^{n}\log\left(\frac{x_i}{\beta}\right)+\frac{n}{\gamma}=0.\end{aligned}$$

From these equations, it is clear that the ML estimator cannot be expressed in closed form and, therefore, the above system must be solved numerically.

We assessed the performance of the ML method via a Monte Carlo simulation study. We first consider the case when one of the parameters is assumed to be known and the aim is to estimate the other two. This is motivated by the fact that in practical applications we found that the subfamily $\mathrm{PM}(1,\beta,\gamma)$ is rich enough to modelling real data sets. From the results obtained, it can be concluded that the ML method provides acceptable estimates of the parameters of the PM family when one of the parameters is assumed to be known.

The parameter estimation problem for the PM family in the general case, that is, when it is assumed that the three parameters are unknown, presents some problems, whose study deserves a deeper study.

4 A real data set application

This section considers a real data set previously studied in Cordeiro and Lemonte [2] on the breaking stress of carbon fibres (in Gba), where the data were fitted by using the Birnbaum-

Saunders (BS) and β-Birnbaum-Saunders (β-BS) distributions. As in [2], we calculated the Akaike information criterion AIC (see Akaike [1]) and the Bayesian information criterion BIC (see Schwarz [7]), which are defined as follows

$$\text{AIC} = 2r - 2\log L, \qquad \text{BIC} = -2\log L + r\log n,$$

where r is the number of parameters and L denotes the maximized value of the likelihood function. In addition to the BS and β-BS distributions, we consider the PM$(1, \beta, \gamma)$ family. Table 1 shows the ML estimated parameters and the AIC and BIC values for each distribution. From these results, it can be concluded that the PM distribution provides a better fit than the BS and β-BS probability models.

Table 1: Breaking stress of carbon fibres: Model, ML estimates, AIC and BIC values.

Model	ML estimates	AIC	BIC
PM$(1, \beta, \gamma)$	$\hat{\beta} = 2.810$, $\hat{\gamma} = 1.394$	176.11	180.49
β-BS(α, β, a, b)	$\hat{\alpha} = 1.044$, $\hat{\beta} = 57.600$, $\hat{a} = 0.193$, $\hat{b} = 1876.732$	190.71	199.47
BS(α, β)	$\hat{\alpha} = 0.437$, $\hat{\beta} = 2.515$	204.38	208.75

Acknowledgements

The research in this paper has been partially funded by the research projects: CTM2015–68276–R of the Spanish Ministry of Economy and Competitiveness (P. Jodrá, M.V. Alba-Fernández), FONDECYT 1130495 (H. Gómez) and MTM2014-55966-P of the Spanish Ministry of Economy and Competitiveness (M.D. Jiménez-Gamero).

References

[1] H. Akaike, *A new look at the statistical model identification*, *IEEE Trans. Automat. Control* **19** (1974) 716–723.

[2] G.M. Cordeiro and A.J. Lemonte, *The β-Birnbaum-Saunders distribution: An improved distribution for fatigue life modeling* Comput. Statist. Data Anal. **55** (2011) 1445–1461.

[3] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey and D.E. Knuth, *On the Lambert W function*, Adv. Comput. Math. **5** (1996) 329–359.

[4] P. Jodrá, M.V. Jiménez-Gamero and M.V. Alba-Fernández, *On the Muth Distribution*, Math. Model. Anal. **20** (2015) 291–310.

[5] M.S. Milgram, *The generalized integro-exponential function*, Math. Comp., **44** (1985) 443–458.

[6] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. URL http://www.R-project.org/.

[7] G. Schwarz, *Estimating the dimension of a model* Ann. Statist., **6** (1978) 461–464.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Dynamics of a Disk on a Rotating Plane with Friction

A. V. Karapetyan[1]

emails: avkarapetyan@yandex.ru

Abstract

The problem of a motion of a circular plate (disk) on a horizontal plane with dry friction uniformly rotating about the vertical is considered. It is shown that the system has two invariant set (attractor and repeller). The general solution of equations of motion of a disk is given in the case of small coefficient of friction.

The considered problem genegalizes problems of the dynamics of a disk on a fixed plane [1] and of a material point on a rotating plane [2].

Key words: friction, disk, rotating plane

The horizontal plane Oxy rotates about the vertical Oz with a constant angular velocity $\boldsymbol{\Omega} = \Omega \boldsymbol{e}_z$ ($\Omega = \text{const}$). A homogeneous disk moves in a plane Oxy under an action of dry friction forces applied to all points of the disk.

Let $\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y$ be a radius-vector of the disk center, $\boldsymbol{u} = \dot{\boldsymbol{r}} = \dot{x}\boldsymbol{e}_x + \dot{y}\boldsymbol{e}_y$ be a relative velocity of this center, $\boldsymbol{\omega} = \omega \boldsymbol{e}_z$ be a relative angular velocity of the disk. Point denotes a derivative by time in the rotating frame $Oxyz$.

The equations of motion of the disk takes the form

$$m\{(\boldsymbol{u} + [\boldsymbol{\Omega}, \boldsymbol{r}])^{\cdot} + [\boldsymbol{\Omega}, \boldsymbol{u} + [\boldsymbol{\Omega}, \boldsymbol{r}]]\} = \boldsymbol{F}, \quad \frac{1}{2} m a^2 \dot{\omega} = M \tag{1}$$

Here m and a are correspondently a mass and a radius of the disk, $\boldsymbol{F}$ is the main vector of the friction forces and M is the vertical component of the main vector of moments of friction forces with respect to the disk center. Accordingly [3]

$$\boldsymbol{F} = -mkp\boldsymbol{u}, \quad M = -makq\omega$$

$$\begin{aligned} p &= \frac{g}{\pi} \int\limits_0^1 s\mathrm{d}s \int\limits_0^{2\pi} \frac{(u - ws\sin\alpha)^2}{(u^2 - 2uws\sin\alpha + w^2 s^2)^{3/2}} \mathrm{d}\alpha = p(u, w), \\ q &= \frac{g}{\pi} \int\limits_0^1 s^3\mathrm{d}s \int\limits_0^{2\pi} \frac{(ws - u\sin\alpha)^2}{(u^2 - 2uws\sin\alpha + w^2 s^2)^{3/2}} \mathrm{d}\alpha = q(u, w) \end{aligned} \tag{2}$$

 ISBN: 978-84-608-6082-2

($u = |\boldsymbol{u}|$, $w = a\omega$, $k > 0$ is the coefficient of Coulomb friction, g is the gravity acceleration)

Thus the equations (1) can be written in the form

$$\ddot{x} - 2\Omega\dot{y} - \Omega^2 x = -kp\dot{x}, \quad \ddot{y} + 2\Omega\dot{x} - \Omega^2 y = -kp\dot{y}, \quad \dot{w} = -2kqw \tag{3}$$

If $\Omega = 0$ then equation (3) describe the dynamics of the disk on a fixed plane and if $a = 0$ (i.e. $w = 0$) the system of the first and second equations (3) describes the dynamics of a material point on the rotating plane.

Equations of motion of the disk (3) permit the invariant sets

$$w = 0 \tag{4}$$

$$r = 0, \quad u = 0 \qquad (r = |\boldsymbol{r}|) \tag{5}$$

The disk dynamics along the set (4) consides with the point dynamics investigated in [2]. In particular, if the disk is at the rest relatively the rotating plane at the initial time $t = 0$ and its center is situated inside the circle of the radius $R = kg/\Omega^2$ (with the center on the axis of rotation of the plane), then the disk will be at the relative rest at all time:

$$r(t) \equiv r_0, \quad u(t) \equiv 0, \quad w(t) \equiv 0 \quad \text{for} \quad r(0) = r_0 < R, \quad u(0) = w(0) = 0.$$

Note that the set (4) is the attractor of the system, because

$$(w^2)^{\cdot} = -4kqw^2 < 0 \quad \forall w \neq 0$$

Let $w(0) = w_0 \neq 0$. Without loss of generality we assume that $w_0 > 0$ (the case $w_0 < 0$ can be considered similary). Then $w(t) \geqslant 0$ $\forall t \geqslant 0$ because (4) is the invariant set. Evidently along the set (5)

$$w(t) = w_0 - \frac{4}{3}kgt \quad (t \in [0, T]) \quad \text{and} \quad w(t) \equiv 0 \quad \forall t \geqslant T = \frac{3w_0}{4kg}$$

Note that the set (4) is the repeller of the system, because

1. $(u^2 - \Omega^2 r^2)^{\cdot} = -2kpu^2 < 0 \quad \forall u \neq 0$,
2. the function $u^2 - \Omega^2 r^2 < 0$ can take negative values in the neibourhood of set (5),
3. $(u^2 - \Omega^2 r^2)^{\cdot} = 0$, if and only if $u = 0$ and (see (3)) $r = 0$.

Thus there exists positive value $\varepsilon > 0$ such that for arbitrary positive value $\delta > 0$ and arbitrary r_0, u_0 and w_0, satisfying the condition

$$u_0^2 + \Omega^2 r_0^2 < \delta, \quad u_0^2 - \Omega^2 r_0^2 < 0, \quad w_0 \neq 0 \tag{6}$$

one can find the time $t > 0$ such that $u^2(t) + \Omega^2 r^2(t) = \varepsilon$.

Note that values $u_0 = 0$, $0 < r_0^2 < \delta/\Omega^2$ satisfy to conditions (6). It means that for zero initial value of a relative velocity of the disk centre ($u_0 = 0$) and arbitrarily small nonzero initial values of a relative angular velocity of the disk ($w_0 = a\omega_0 \neq 0$) disk begins to slide for arbitrarily small nonzero distance between the disk center and the rotation axis of rotation of the plane ($r_0 \neq 0$).

Let us assume that a coefficient of Coulomb friction is very small ($0 < k << 1$) and find the motion of the disk for arbitrary initial conditions outside invariant sets (4) and (5), i.e. for $\Omega^2 r_0^2 + u_0^2 > 0$, $w_0 \neq 0$ (without loss of generality we assume that $w_0 > 0$).

Introduce a complex variable $z = x + iy$ ($i^2 = -1$) and present the system (3) in the form

$$\ddot{z} + 2i\Omega\dot{z} - \Omega^2 z = -kp(|\dot{z}|, w)\dot{z}, \quad \dot{w} = -2kq(|\dot{z}|, w)w \tag{7}$$

and its solution in the form

$$z(t) = z_0(t) + kz_1(t) + \ldots, \quad w(t) = w_0(t) + kw_1(t) + \ldots \tag{8}$$

Assume that $z(0) = z_0(0)$, $\dot{z}(0) = \dot{z}_0(0)$, $w(0) = w_0(0)$:

$$z_0(0) = z_0, \quad \dot{z}_0(0) = \dot{z}_0, \quad w(0) = w_0 > 0 \quad (|\Omega||z_0| + |\dot{z}_0| \neq 0) \tag{9}$$

$$z_j(0) = 0, \quad \dot{z}_j(0) = 0, \quad w_j(0) = 0 \quad (j = 1, 2, \ldots) \tag{10}$$

Thus zero-order approximation $z_0(t)$, $w_0(t)$ of the solution (8) satisfy to the linear homogeneous system with constant coefficients

$$\ddot{z}_0(t) + 2i\Omega\dot{z}_0(t) - \Omega^2 z_0(t) = 0, \quad \dot{w}_0(t) = 0 \tag{11}$$

and initial conditions (9), and functions $z_j(t)$, $w_j(t)$ ($j = 1, 2, \ldots$) satisfy to nonhomogeneous system

$$\ddot{z}_j + 2i\Omega\dot{z}_j - \Omega^2 z_j = -p_j(t), \quad \dot{w}_j = -2q_j(t) \quad (j = 1, 2, \ldots) \tag{12}$$

and zero initial conditions (10). Here functions $p_j(t)$ and $q_j(t)$ are the coefficient at k^{j-1} in expanded in series function

$$p(|\dot{z}^{(j-1)}(t)|, w^{(j-1)}(t))\dot{z}^{(j-1)}(t), \quad \text{and} \quad q(|z^{(j-1)}(t)|, w^{(j-1)}(t))w^{(j-1)}(t)$$

where $z^{(n)}(t) = \sum_{s=0}^{n} z_s(t)k^s$, $w^{(n)}(t) = \sum_{s=0}^{n} w_s(t)k^s$.

The solution of the Cauchy problem (9), (11) has the form

$$z_0(t) = (z_0 + (\dot{z}_0 + i\Omega z_0)t)e^{-i\Omega t}, \quad w_0(t) \equiv w_0 \tag{13}$$

 ISBN: 978-84-608-6082-2

and the solution of the Cauchy problem (10), (12) has the form

$$z_j(t) = -e^{-i\Omega t}\int_0^t \mathrm{d}\tau \int_0^\tau p_j(\sigma)e^{i\Omega\sigma}\mathrm{d}\sigma, \quad w_j = -2\int_0^t q_j(\tau)\mathrm{d}\tau \quad (j=1,\ldots) \tag{14}$$

Thus the general solution of equations (7) of motion of the disk on a rotating plane with dry friction (in the case of a small coefficient of friction) can be presented in the series (8). The first terms of these series are the explicit function of time (13) and all other terms are the quadratures (14). Note that the series (8) with coefficient (13),(14) are correct only for rotating disk ($w(t) \neq 0$).

If there exist time $T > 0$ such that $w(T) = 0$ then these series are correct only for $t \in [0, T]$. For $t \geqslant T$ the disk moves as a material point. Evidently the series (8) permit to construct the series

$$r(t) = r_0(t) + kr_1(t) + \ldots, \quad u(t) = u_0(t) + ku_1(t) + \ldots \tag{15}$$

Let us investigate the disk dynamics in the first approximation

$$z(t) = z^{(1)}(t) + o(k), \quad w(t) = w^{(1)}(t) + o(k)$$
$$(z^{(1)}(t) = z_0(t) + kz_1(t), \quad w^{(1)}(t) = w_0(t) + kw_1(t))$$
$$r(t) = r^{(1)}(t) + o(k), \quad u(t) = u^{(1)}(t) + o(k)$$
$$(r^{(1)}(t) = r_0(t) + kr_1(t), \quad u^{(1)}(t) = u_0(t) + ku_1(t))$$

Here $r_0(t) = |z_0(t)|$, $u_0(t) = |\dot{z}_0(t)|$, $r_1(t) = \mathrm{Re}(z_0(t)\overline{z_1}(t))/r_0(t)$, $u_1(t) = \mathrm{Re}(\dot{z}_0(t)\dot{\overline{z_1}}(t))/u_0(t)$.
Let us assume (for simplicity) that $z_0 = r_0 > 0$, $\dot{z}_0 = 0$. Then

$$z_0(t) = r_0(1 + i\Omega t)e^{-i\Omega t}, \quad w_0(t) \equiv w_0, \qquad r_0(t) = r_0(1 + \Omega^2 t^2)^{\frac{1}{2}}, \quad u_0(t) = r_0\Omega^2 t$$

Thus $p_1(t) = r_0 f(t)e^{-i\Omega t}$, $\quad q_1(t) = w_0 h(t)$, where

$$f(t) = \frac{g}{\pi}\Omega^2 t\left[\int_0^1 s\mathrm{d}s\int_0^{2\pi}\frac{(r_0\Omega^2 t - w_0 s\sin\alpha)^2}{(r_0^2\Omega^4 t^2 - 2r_0\Omega^2 t w_0 s\sin\alpha + w_0^2 s^2)^{3/2}}\mathrm{d}\alpha\right]$$
$$h(t) = \frac{g}{\pi}\left[\int_0^1 s^3\mathrm{d}s\int_0^{2\pi}\frac{(w_0 s - r_0\Omega^2 t\sin\alpha)^2}{(r_0^2\Omega^4 t^2 - 2r_0\Omega^2 t w_0 s\sin\alpha + w_0^2 s^2)^{3/2}}\mathrm{d}\alpha\right]$$

So

$$z_1(t) = -r_0 l(t)e^{-i\Omega t}, \quad w_1(t) = -2w_0\int_0^t h(\tau)\mathrm{d}\tau, \quad u_1(t) = -r_0\int_0^t f(\tau)\mathrm{d}\tau$$

$$r_1(t) = -r_0 l(t)(1 + \Omega^2 t^2)^{-\frac{1}{2}}, \quad \left(l(t) = \int_0^t \mathrm{d}\tau\int_0^\tau f(\sigma)\mathrm{d}\sigma\right)$$

 ISBN: 978-84-608-6082-2

Finaly

$$\begin{gathered} r^{(1)}(t) = r_0(1+\Omega^2 t^2)^{\frac{1}{2}}\left[1 - k(1+\Omega^2 t^2)^{-1}\int_0^t \mathrm{d}\tau \int_0^\tau f(\sigma)\mathrm{d}\sigma\right] \\ u^{(1)}(t) = r_0\left[\Omega^2 t - k\int_0^t f(\tau)\mathrm{d}\tau\right], \quad w^{(1)}(t) = w_0\left[1 - 2k\int_0^t h(\tau)\mathrm{d}\tau\right] \end{gathered} \tag{17}$$

Note that

$$h(t) > 0, \quad \lim_{t\to+\infty} th(t) = \frac{g}{4r_0\Omega^2} = c$$

i.e. asymptotic of the function $h(t)$ for $t \to +\infty$ is defined by the function c/t. So

$$\lim_{r\to+\infty}\int_0^t h(\tau)d\tau = +\infty$$

i.e. there exists $T > 0$ such that $2k\int_0^t h(\tau)\tau = 1$. Thus $w^{(1)}(T) = 0$. Taking into account that

$$f(0) = 0, \quad \lim_{t\to+\infty} f(t) = \frac{g}{r_0}$$

one can conclude that $f(t)$ is a bounded function. Thus $r^{(1)}(T) > 0$ and $u^{(1)}(T) > 0$ for small k and formula (16) are correct for $t \in [0, T]$. For $t \geqslant T$ the relative angular velocity of the disk equal to zero (in the first approximation) identically and the disk dynamics coincides with the material point dynamics.

This work is supported by Russian Foundation for Basic Research (projects 16-01-00338).

References

[1] A.Yu. Ishlinskii, B.N. Sokolov, F.L. Chernousko. On the motion of planar bodies in the presence of dry friction. Izv. AN SSSR MTT(Mechanics of solids). 4. P. 17-28. 1981.

[2] A.I. Grudev, A.Yu. Ishlinskii, F.L. Chernousko. On the motion of a particle over a rough rotating plane. J. of Appl. Mech. (PMM) V. 53. . 3. P. 281- 288. 1989

[3] A.V. Karapetyan, A.M. Rusinova. A qualitative analysis of the disk on an inclined plane with friction. J. of Appl. Mech. (PMM) V. 75. . 5. P.511- 516. 2011

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The Dynamics of a Heavy Ridid Ellipsoid on a Horizontal Plane with Friction

A. V. Karapetyan[1] **and M. A. Munitsyna**[2]

[1] *Lomonosov Moscow State University,*

[2] *Moscow Institute of Physics and Technology (State University),*

emails: avkarapetyan@yandex.ru, munitsyna@gmail.com

Abstract

The dynamics of an ellipsoid of revolution with a displaced mass centre on a horizontal plane with friction is considered. It is assumed that the mass centre of the ellipsoid lies on its dynamic symmetry axis. The steady motions of the ellipsoid are investigated using the general theory of invariant sets of mechanical systems with symmetry, and a geometrical interpretation of the results is given using generalized Smale diagrams. In the special case when the equatorial and axial radii of the ellipsoid are equal, it is the simplest model of a tippy-top, the localized analysis of the dynamics of which [1,2] and a global qualitative analysis [3,4] have been given earlier.

Key words: dissipative mechanical systems with symmetry, Smale diagrams.

Consider a heavy non-uniform dynamically symmetrical ellipsoid or revolution of mass m on a horizontal plane. Suppose a and c are the equatorial and axial radii of the ellipsoid and is the distance between its geometrical centre O and its mass centre S.It is assumed that the mass centre of the ellipsoid is on its axis of symmetry and the equation of the surface of the ellipsoid in its principal central axes

$$f(\boldsymbol{x}) = \frac{x_1^2 + x_2^2}{a^2} + \frac{(x_3 - s)^2}{c^2} - 1 = 0$$

We will introduce the following variables: $\boldsymbol{v}$ is the velocity of the mass centre of the ellipsoid, $\boldsymbol{\omega}$ is its angular velocity and $\boldsymbol{\gamma}$ is the unit vector of the rising vertical. The sliding velocity of the ellipsoid is given by the relation $\boldsymbol{u} = \boldsymbol{v} + [\boldsymbol{\omega}, \boldsymbol{r}]$, where $\boldsymbol{r}$ is the radius

 ISBN: 978-84-608-6082-2

vector of the point of tangency of the ellipsoid with the plane, defined by the equation $\boldsymbol{\gamma} = -f(\boldsymbol{r})/|f(\boldsymbol{r})|$. We will write the components of the radius vector in the system $Sx_1x_2x_3$

$$r_1 = -a^2\gamma_1/\rho, \quad r_2 = -a^2\gamma_2/\rho, \quad r_3 = s - c^2\gamma_3/\rho, \quad \rho = ((c^2 - a^2)\gamma_3^3 + a^2)^{1/2}$$

A gravitational force $\boldsymbol{P} = -mg\boldsymbol{\gamma}$, the normal component of the reaction of the support plane $\boldsymbol{N} = N\boldsymbol{\gamma}$ and the sliding friction force $\boldsymbol{F}$ act on the ellipsoid. The equations of the motion of the ellipsoid, referred the principal central axes of inertia have the form

$$m\dot{\boldsymbol{v}} + [\boldsymbol{\omega}, m\boldsymbol{v}] = (N - mg)\boldsymbol{\gamma} + \boldsymbol{F} \tag{1}$$

$$\dot{\boldsymbol{\omega}} + [\boldsymbol{\omega}, \boldsymbol{\omega}] = [\boldsymbol{r}, N\boldsymbol{\gamma} + \boldsymbol{F}] \tag{2}$$

$$\dot{\boldsymbol{\gamma}} + [\boldsymbol{\omega}, \boldsymbol{\gamma}] = 0 \tag{3}$$

$$(\boldsymbol{v} + [\boldsymbol{\omega}, \boldsymbol{r}], \boldsymbol{\gamma}) = 0 \tag{4}$$

where $= (A, A, C)$ is the central inertia tensor or the ellipsoid. Equation (1) expresses the theorem of the change in the momentum of the ellipsoid, (2) expresses the theorem of the change in the angular momentum about the mass centre, (3) expresses the condition for the vector $\boldsymbol{\gamma}$ to be constant in an absolute frame of reference, and (4) expresses the condition for the ellipsoid to be in contact with the supporting plane. If the friction force is specified in the form $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{v}, \boldsymbol{\omega}, \boldsymbol{\gamma}, N)$, system (1)–(4) is closed with respect to the variables $\boldsymbol{v}$, $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and N.

We will assume that the sliding friction force satisfies the natural conditions

$$(\boldsymbol{F}, \boldsymbol{u}) < 0 \qquad \boldsymbol{u} \neq \boldsymbol{0}, \qquad \boldsymbol{F} = \boldsymbol{0} \qquad \boldsymbol{u} = \boldsymbol{0}, \tag{5}$$

which the classical model of viscous friction or the generalized model of dry friction [5-7] , for example, satisfy.

System (1)–(4) allows of the energy relation $\dot{H} = (\boldsymbol{F}, \boldsymbol{u}) \leq 0$,

$$H = \frac{1}{2}mv^2 + \frac{1}{2}(J\boldsymbol{\omega}, \boldsymbol{\omega}) - mg(\boldsymbol{r}, \boldsymbol{\gamma}) - mgc \tag{6}$$

is the total mechanical energy. Moreover, the change in the quantity $K = (\boldsymbol{\omega}, \boldsymbol{r})$ in the solution of the system has the form

$$\dot{K} = (\boldsymbol{\omega}, \dot{r} + [\boldsymbol{\omega}, \boldsymbol{r}]) = \frac{a^2}{\rho^3}(c^2 - a^2)(\omega_2\gamma_1 - \omega_1\gamma_2)\left[(\mathbb{J}\boldsymbol{\omega}, \boldsymbol{\gamma})\gamma_3 - C\omega_3\right] \tag{7}$$

Note that, in the case of a sphere ($a = c$) a Jellett integral occurs and the quantity does not change.

One can show that there is no slippage only on motion

$$\gamma_1 = \gamma_2 = 0, \ \gamma_3 = \pm 1, \quad \omega_1 = \omega_2 = 0, \ \omega_3 = \omega =, \quad N = mg \tag{8}$$

 ISBN: 978-84-608-6082-2

$$\begin{gathered}\gamma_1 = \sqrt{1-\gamma_3^2}\cos\Omega t,\ \gamma_2 = \sqrt{1-\gamma_3^2}\sin\Omega t,\\ \omega_1 = \omega_0\gamma_1,\ \omega_2 = \omega_0\gamma_2,\ \omega_3 = \omega_0\gamma_3 + \Omega, \qquad N = mg,\end{gathered} \tag{9}$$

Hence [8], the set of all steady motions can be represented in the (k^2, h),plane, where k and h are the initial values of the total mechanical energy H and the quantity K respectively, in the form of $h = h(k^2)$, curves, all the points of which, in view of the above, are invariant with respect to the phase flow of the system considered. Then all the remaining points in the (k^2, h)plane move in the direction of decreasing h.

The curves $h_\pm(k^2)$, , corresponding to uniform rotations, are written explicitly as follows:

$$h_\pm = \frac{1}{2}\frac{k^2}{C(c \mp s)^2} \mp mgs \tag{10}$$

while the curves $h_*(k^2)$, , corresponding to regular precessions, are specified parametrically in the form

$$\begin{aligned}h_* &= mg\Big(-c - s\gamma_3 + \rho - \tfrac{1}{2}\left(A(1-\gamma_3^2) + C\tfrac{(c^2\gamma_3 - s\rho)^2}{a^4}\right)\tfrac{s-(c^2-a^2)\gamma_3/\rho}{C(c^2\gamma_3 - s\rho)/a^2 - A\gamma_3}\Big)\\ k_*^2 &= -mg\tfrac{[Aa^4(1-\gamma_3) + C(c^2\gamma_3 - s\rho)^2]^2}{a^2\rho^2}\tfrac{(s-(c^2-a^2)\gamma_3/\rho)}{C(c^2\gamma_3 - s\rho)/a^2 - A\gamma_3}\end{aligned} \tag{11}$$

We recall that, in the case of sphere, a Jellett integral occurs and the value of K does not change, and hence point in the (k^2, h), plane, which are not on the steady motion curves, move in the direction of decreasing $h$along the corresponding axis. A complete parametric analysis of the problem has been carried out in the case [1,4]: the parameter plane of the sphere is divided into seven regions, each of which has its own generalized Smale diagram. In the case of an ellipsoid, close to a sphere, i.e. when $c/a = 1 + \varepsilon$, $\varepsilon << 1$,the relation $\dot{K} << \varepsilon$,holds, i.e. points in the (k^2, h), plane, not on the steady motion curves, move in the direction of decreasing h inside a narrow strip, parallel to this axis.

In Fig.1 we show numerically constructed trajectories of the points (the dark curves). The initial position of the axis of symmetry of the ellipsoid is almost vertical ($\gamma_3(0) = 0.9$), the mass centre is fixed and is below the geometrical centre of the ellipsoid, while the angular velocity is directed vertically upwards. Its value in experiments 1-5 vary uniformly from 20 to 100 $\sec^{-1}$.

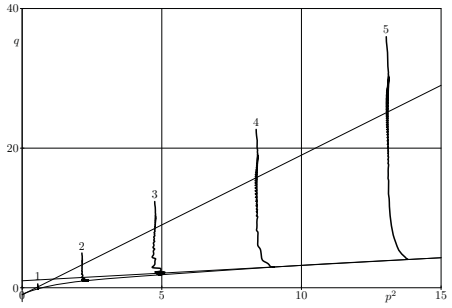


Fig. 1

 ISBN: 978-84-608-6082-2

The abscissa axis corresponds to the square of the dimensionless value of the quantity K ($p^2 = k^2/(Cmgs)$), while the ordinate axis corresponds to the dimensionless value of the total mechanical energy $q = h/(mgs)$. In Fig.1 we also show the straight lines (10), corresponding to rotations (the straight line passing through the point $(0, -1)$ corresponds to rotations with the lowest position of the mass centre, while the straight line passing through the point $(0, 1)$ corresponds to rotations with the highest position of the mass centre), and their connecting precession curve (11). Hence, in experiments 1-3 the final motion of the ellipsoid is a precession, while in experiments 4 and 5 the final motions are uniform rotations around the vertical, for which the mass centre of the ellipsoid is situated above its geometrical centre.

This work is supported by Russian Foundation for Basic Research (projects 16-01-00338, 14-01-00432).

References

[1] *Contensou P.* Couplage entre frottemeuz et frattement de pivotement dans la theoric de la toupee // Kreisel-probleme. Derlin, 1963. .

[2] *Magnus, Kreisel.* Theorie und Anwendungen. Berlin: Springer, 1971

[3] *Karapetyan A.V.* A global qualitative analysis of the dynamics of the Chinese top (tippy-top), Izv Akad Nauk MTT 3,33-41,2008.

[4] *Karapetyan A.V.* Invariant sets of mechanical systems with symmetry. Problems of Stability and Control. Collected Papers Commemorating the 80th Birthday of V.M.Matrosov. Moscow: Fizmatlit; 2013.

[5] *Zuravlev V.F.* The model of dry friction in the problem of the rolling of a rigid body, JAMM 1998; 62 (5): 705-710.

[6] *Karapetyan A.V.* A two-parameter model of friction. JAMM 2009; 73 (4) : 515-519.

[7] *Munitsyna M.A.* A model of friction in the case of plane elliptical contact of a body with a support plane, Nelin Dinamika 2012; 8 (4) : 705-712.

[8] *Karapetyan A.V.* Generalized Smale diagrams and their application to problems of the dynamics of systems with friction. Analytical Mechanics, Stability and Control. Proceeding of the 10th International Chetayev Conference. Vol. 1 Section 1. Analytical Mechanics. Kazan, 2012. Kazan : Izd. Kazan. Gos. Univ; 2012: 247-258.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Persistence analysis of the age-structured population model on several patches

Vladimir Kozlov[1], Sonja Radosavljevic[1], Vladimir Tkachev[1] and Uno Wennergren[2]

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

[2] ***Department of Physics, Chemistry, and Biology, Linköping University*

emails: vladimir.kozlov@liu.se, sonja.radosavljevic@liu.se, vladimir.tkatjev@liu.se, uno.wennergren@liu.se

Abstract

We consider a system of nonlinear partial differential equations that describes an age-structured population living in changing environment on N patches. We prove existence and uniqueness of solution and analyze large time behavior of the system in time-independent case, for periodically changing and for irregularly varying environment. Under the assumption that every patch can be reached from every other patch, directly or through several intermediary patches, and that net reproductive operator has spectral radius larger than one, we prove that population is persistent on all patches. If the spectral radius is less or equal one, extinction on all patches is imminent.

Key words: Age-structured population model, temporal variability, spatial heterogeneity, density-dependency, dispersion, large time behavior, periodic solutions
MSC 2000: AMS codes: 47N60, 47B65, 62P10

1 Introduction

Various natural processes and human activities are responsible for changes in habitat structure, quality and its fragmentation. A consequence of habitat reduction or destruction can be very severe and result in extinction of populations and loss of biodiversity. Understanding how habitat fragmentation and heterogeneity, climate change and seasonality interact with internal factors of population growth and influence population dynamics may provide a useful tool for conservation and managing of species ([13]).

The idea that spatial structure is important and should be considered in studying population dynamics is not new. It is clear that there are no two identical habitats and that some are better than others. This can be linked to habitat heterogeneity and to the source-sink systems. A high quality habitat (source) yields positive growth rate of population, while low quality habitat (sink) has the opposite effect. For a species that inhabits several patches, possibility to move from one patch to another can be crucial for survival. Migration from source to sink has been studied and its stabilizing effect for population or even ecosystem is well-known. The importance of migration is in the fact that arrival of immigrants to the sink can save local population from extinction ([2], [8]).

No matter how important, spatial structure is not the only factor that influence population dynamics. It should be viewed as one of the pivotal factors and studied in combination with population structure, temporal variability and density-dependence (see for example [3]). All mentioned factors have relative importance for population dynamics which differs for terrestrial and marine species, vertebrate and invertebrate, plants and animals, large and small populations etc. The challenge is to make a model that includes all these factors and yet remains simple enough in order to be applicable to real world problems.

However, in spite of the evidence that various internal and external factors interact and contribute to population dynamics and complexity of ecological systems, a usual practice is to study only one or few of them, but not all of them simultaneously. We mention some of them. Gurtin and McCamy introduced age-structured density dependent model in [11]. Chipot considered age-structured models with time and density-dependency in [6] and [7]. Allen in [1] studies different dispersion mechanisms for the logistic model without age-structure and without time-dependency in vital rates. Cui and Chen in [4] and [5] study the effect of diffusion on a single species or for predator-prey system, but they do not consider structured population. The system of delay differential equations developed by So, Wu and Zou in [16] deals with population divided into two age classes (immature and adult) and assumes that population inhabits two identical patches and that the vital rates are time-independent. In [21], Weng, Xiao and Zou extended the model of So, Wu and Zou in [16] and considered dynamics of population on three patches. Terry in [20] investigates population of two stages and on two patches and discusses persistence of population for different birth functions and dispersion patterns. Takeuchi in [18] and [19] examined stability and effect of diffusion of generalized Lotka-Volterra systems. Hastings discussed stabilizing effect of dispersal in [12].

In contrast to the previous papers, we consider age-structure, time and density-dependence and dispersion between patches and investigate their impact on population dynamics. In formulation of the model we rely on several assumptions: population is age-structured and inhabits N patches. In addition to this, environment changes with time, causing the change in the vital rates, competition level and dispersion coefficients. Competition for the resources occurs within each age-class which results in increased density-dependent

mortality. The patches are not identical and they do not have to be physically close, because many pest species exploit trading and transportation networks to move from one patch to another.

2 Formulation of the model

Under the made assumptions on mind, the system of balance equations for the model becomes:

$$\frac{\partial n_k(a,t)}{\partial t} + \frac{\partial n_k(a,t)}{\partial a} = -\mu_k(a,t)n_k(a,t)\left(1 + \frac{n_k(a,t)}{L_k(a,t)}\right) + \sum_{j=1}^{N} D_{kj}(a,t)n_j(a,t), \quad 1 \le k \le N, \tag{1}$$

and the boundary and initial conditions are:

$$n_k(0,t) = \int_0^\infty m_k(a,t)n_k(a,t)\,da, \quad t > 0, \tag{2}$$

and

$$n_k(a,0) = f_k(a), \quad a > 0, \tag{3}$$

where $n_k(a,t)$ is the number of individuals of age a at time t on patch k, $\mu_k(a,t)$ is the death rate, $L_k(a,t)$ the regulating function,

$$D = \|D_{kj}(a,t)\|_{1\le k,j\le N}$$

the matrix of dispersion coefficients, $m_k(a,t)$ the birth rate and $f_k(a)$ the initial distribution of population.

The regulating function $L_k(a,t)$ represents limitations imposed on individuals by environment (or available resource per capita) and the logistic term $\frac{\mu_k(a,t)n_k^2(a,t)}{L_k(a,t)}$ describes increment in mortality due to population density. The regulation function resembles carrying capacity because for small values of $L_k(a,t)$ mortality grows and for its large values, logistic term tends to zero. Unlike carrying capacity it does not represent maximal supported population. We will use it as a starting point in study of relation between age-structure and density-dependence.

Underlying assumption is that population does not have to occupy all patches at initial time, but in order for population to survive, sufficiently young individuals must occupy at least one patch. Nonnegative dispersion coefficients $D_{kj}(a,t)$ for $k \neq j$ define proportion of individuals of age a at time t on patch j that migrates to patch k. Every patch k can be reached from every other patch j, possibly via several passing patches. Dispersion

coefficients $D_{kk}(a,t) \leq 0$ define proportion of population of age a at time t that leaves patch k.

Furthermore, we also assume that the dispersion matrix is *essentially positive*, i.e. for any $k \neq j$ there exist pairwise distinct indices $m_0, m_1, \ldots, m_s$ such that $m_0 = k$, $m_s = j$ and $D_{m_i m_{i+1}}(a,t) > 0$ for any $0 \leq i \leq s-1$. The latter condition has the following natural explanation. Let us associate a directed graph $\Gamma(D,x)$ to the dispersion matrix D as follows: the nodes of the graph are presented by the N patches and the edges represent the transitions between patches according to the dispersion coefficients $D_{kj}(x)$ such that the (k,j)-entry of the incidence matrix of $\Gamma(D,x)$ is 0 if $D_{kj}(x) = 0$, and 1 if $D_{kj}(x) > 0$. Then D satisfy the above condition if and only if the graph $\Gamma(d,x)$ is connected (i.e. there is a path between every pair of vertices). An examples of an essentially positive matrix is

$$\begin{pmatrix} * & + & 0 \\ 0 & * & + \\ + & 0 & * \end{pmatrix}$$

where $*$ means an entry of an unspecified sign.

3 The main results

In this section we briefly describe our strategy and the principal results. The first main result is the existence and uniqueness of solution to the problem (1)–(3) in the class of bounded continuous functions. To this end, we consider the problem with time-independent vital rates, regulating function and dispersion coefficients. Namely, in order to determine the number of newborns

$$\rho_k(t) := n_k(0,t), \quad t \geq 0, \quad 1 \leq k \leq N,$$

we introduce two auxiliary initial value problems as follows. Let $\Phi(x,y;\rho)$ and $\Psi(x,y;f)$ denote respectively the solutions of the following initial value problems:

$$\frac{dh_k(x)}{dx} = -\mu_k(x,x+y)\left(h_k(x) + \frac{h_k^2(x)}{L_k(x,x+y)}\right) + \sum_{j=1}^{N} D_{kj}(x,x+y)h_j(x) = 0, \quad h(0) = \rho(y)$$

$$\frac{dh_k(x)}{dx} = -\mu_k(x+y,x)\left(h_k(x) + \frac{h_k^2(x)}{L_k(x+y,x)}\right) - \sum_{j=1}^{N} D_{kj}(x+y,x)h_j(x), = 0, \quad h(0) = f(y).$$

It can be shown that each of the latter problems has a unique *nonnegative* solution.

Then, the original problem can be reduced to the integral equation

$$\rho(t) = \mathcal{K}\rho(t) + \mathcal{F}f(t), \tag{4}$$

where $\rho(t) = (\rho_1(t), \dots, \rho_N(t))$ and the operators $\mathcal{K}$ and $\mathcal{F}$ are defined componentwise by

$$(\mathcal{K}\rho)_k(t) = \int_0^t m_k(a,t)\Phi_k(a, t-a; \rho)\, da \tag{5}$$

$$(\mathcal{F}f)_k(t) = \int_t^\infty m_k(a,t)\Psi_k(a, a-t; f)\, da, \tag{6}$$

In the time-independent case, an important role in description and analysis of solutions to (4) plays the so-called characteristic equation, i.e. the (unique) solution to

$$\rho = \bar{\mathcal{K}}\rho. \tag{7}$$

Here the operator $\bar{\mathcal{K}}$ is given by

$$(\bar{\mathcal{K}}\rho)_k := \int_0^\infty m_k(a)\Phi_k(a; \rho)\, da, \qquad \rho \in \mathbb{R}^N_+,$$

i.e. when the newborns function ρ is constant for all $t \geq 0$. Clearly, $\rho = 0$ is always a solution of the characteristic equation, but our goal is to find out when a nontrivial solution $\rho \in \mathbb{R}^N_+ \setminus \{0\}$ does exist and is unique.

An important ingredient of our approach is the operator $\mathscr{R}_0$ defined by the right hand side in (2), where $n_k(a,t)$ solves (1) but without nonlinear term (formally assuming that $\frac{1}{L_k(a,t)} \equiv 0$) and with the boundary condition $n_k(0,t) = \rho_k = const$. We show that it can be alternatively defined as the blow-up of $\bar{\mathcal{K}}$:

$$\mathscr{R}_0\rho = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\bar{\mathcal{K}}\varepsilon\rho.$$

Due to assumptions on dispersion coefficients $D_{kj}(a,t)$, this operator can be shown to be strongly positive and linear. Then an important corollary of the Krein-Rutman theorem is that its spectral radius $\sigma(\mathscr{R}_0)$ is equal to the largest positive eigenvalue.

Thus defined operator $\mathscr{R}_0$ is called the *net reproductive map* and $\sigma(\mathscr{R}_0)$ is the net reproductive rate. Notice that in the one-dimensional case, $\sigma(\mathscr{R}_0)$ coincides with the net reproductive rate R_0 which is a well-established concept and it is defined by

$$R_0 = \int_0^\infty m(a)e^{-\int_0^a \mu(v)dv}\, da.$$

The net reproductive operator is just multiplication by this number. In [22] we proved that R_0 corresponds to the characteristic equation

$$\int_0^\infty \frac{m(a)e^{-\int_0^a \mu(v)dv}}{1 + n^*(1 - e^{-\int_0^a \mu(v)dv})}\, da = 1,$$

 ISBN: 978-84-608-6082-2

through which it is related to the average number of newborns n^*. Moreover, population declines for $R_0 \leq 1$ and grows for $R_0 > 1$.

We show that our definition is in consistence with the classical definition in one-dimensional case. Within this framework, we are able to characterize stationary solutions to the characteristic equation (7). Namely, the following dichotomy holds

Theorem 1. *If $\sigma(\mathscr{R}_0) \leq 1$, then the characteristic equation has only trivial solution $\rho^* = 0$. If $\sigma(\mathscr{R}_0) > 1$, then the characteristic equation has exactly one nontrivial solution ρ^* whose all components are positive.*

One of ingredients of the proof is the following monotonicity result which has a considerable interest in itself. Here and in what follows we use the standard notation vector order relation: given two vectors $x, y \in \mathbb{R}^N$ one defines $x \leq y$ if $x_i \leq y_i$ for all $1 \leq i \leq n$. Further, $x < y$ if $x \leq y$ and $x \neq y$, and $x \ll y$ if $x_i < y_i$ for all $1 \leq i \leq n$.

Lemma 2. *Let $w(x) = (w_1, ..., w_N)$ be locally Lipschitz functions satisfying $w(0) \geq 0$ and*

$$w_k'(x) \geq \sum_{j=1}^{N} d_{kj}(x) w_j(x), \quad 1 \leq k \leq N, \quad x \in [0, b), \tag{8}$$

where $d_{kj}(x)$ are continuous in $[0, b)$ and $d_{kj}(x) \geq 0$ for all $j \neq k$. Then $w(x) \geq 0$ on $[0, b)$. Furthermore, if $(d_{kj})_{1 \leq k,j \leq N}$ is essentially positive and $w(0) > 0$ then $w(x) \gg 0$ on $(0, b)$.

Another important ingredient of the proof is a suitable generalization of the existence and the uniqueness results for monotone and concave operators established by Krasnoselskii and Zabreiko in [14].

Our next result concerns solutions of the integral equation (4). We show that an a priori nonconstant solutions tends to the (constant) solution of the characteristic equation (7). In other words,

Theorem 3. *Let $\rho = \rho(t)$ be a solution to (4). If $\sigma(\mathscr{R}_0) \leq 1$ then $\rho(t) \to 0$ as $t \to \infty$. If $\sigma(\mathscr{R}_0) > 1$, then $\rho(t) \to \rho^*$ as $t \to \infty$, where ρ^* is defined by Theorem 1. In particular,*

$$\text{if } \sigma(\mathscr{R}_0) \leq 1, \text{ then } N(t) \to 0 \text{ as } t \to \infty,$$

and

$$\text{if } \sigma(\mathscr{R}_0) > 1, \text{ then } N(t) \to \int_0^\infty \Phi(a; \rho^*)\, da \text{ as } t \to \infty,$$

where $N(t) = \int_0^\infty n(a, t)\, da$ is the total population and $\Phi(a; \rho^*)$ is a solution to the original problem with ρ^* as the initial condition.

In this way, the net reproductive rate $\sigma(\mathscr{R}_0)$ effectively determines large time behavior of population on N patches in constant environment. Here, as in the one-dimensional case, $\sigma(\mathscr{R}_0) \leq 1$ implies extinction of population on all patches, while $\sigma(\mathscr{R}_0) > 1$ grants persistence of population.

4 Periodically varying environment

Most natural habitats are positively autocorrelated, see for example [17]. Thus, the assumption that the vital rates, regulating function and dispersal coefficient are changing periodically with respect to time is reasonable. In studying large time behavior of solution to equation (4) in periodically changing environment, a pivotal role belongs to the characteristic equation $\rho(t) = \widetilde{\mathcal{K}}\rho(t)$, where operator $\widetilde{\mathcal{K}}$ is given by the right hand side of (2) and $n_k(a,t)$ solves (1) with periodic boundary condition $n_k(0,t) = \rho_k(t)$, $t > 0$.

Similarly to the previous situation, we introduce the net reproductive operator $\widetilde{\mathscr{R}}_0$ by the right hand side of (2) assuming that $n_k(a,t)$ solves (1) without nonlinear term and with periodic boundary condition $n_k(0,t) = \rho_k(t)$. In this case, we have that

$$\widetilde{\mathscr{R}}_0\rho = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\widetilde{\mathcal{K}}\varepsilon\rho.$$

Operator $\widetilde{\mathscr{R}}_0$ is strictly positive, linear and defined on space of periodic continuous functions. Its spectral radius $\sigma(\widetilde{\mathscr{R}}_0)$ is equal to the largest eigenvalue and it is called the net reproductive rate.

One of the main results about periodic case is that if $\sigma(\widetilde{\mathscr{R}}_0) \le 1$, then the characteristic equation $\rho(t) = \widetilde{\mathcal{K}}\rho(t)$ has only trivial solution $\rho \equiv 0$. If $\sigma(\widetilde{\mathscr{R}}_0) > 1$, then the characteristic equation has exactly one nontrivial solution $\rho^*(t)$, where all components are positive periodic functions. Furthermore, we show that the number of newborns in periodic case converges to a unique periodic solution of the characteristic equation. For the total population, the result can be formulated in the following way:

$$\text{if } \sigma(\widetilde{\mathscr{R}}_0) \le 1, \text{ then } N(t) \to 0 \text{ as } t \to \infty,$$

and

$$\text{if } \sigma(\widetilde{\mathscr{R}}_0) > 1, \text{ then } N(t) \to \textstyle\int_0^\infty \Phi(a, t-a; \rho^*)\, da \text{ as } t \to \infty,$$

where $\Phi(a,t;\rho^*)$ is a solution to the original initial value problem and $\rho^*(t)$ is a periodic solution to the characteristic equation. In this situation, as in the time-independent case, the net reproductive rate $\sigma(\widetilde{\mathscr{R}}_0)$ determines extinction or persistence of population on all patches.

Upper and lower bounds for population growth can be found even if environment is changing irregularly. Namely, if the parameters of the original model are estimated from above and below by periodic functions for large time, then two periodic problems can be formulated. One of these periodic problems is an upper bound for the original problem, and the other is a lower bound. The number of newborns in the original problem, $\rho(t)$, is then estimated by the number of newborns in periodic problems. In other words, the following holds:

$$\rho^-(t) - \varepsilon \le \rho(t) \le \rho^+(t) + \varepsilon,$$

where $\rho^{\pm}(t)$ are solutions of the characteristic equations for periodic problems and $\varepsilon > 0$ is an arbitrary small number.

In order to explain the influence of dispersion on persistence of population, we compare the system of N patches with dispersion with the system of N isolated patches. To obtain the net reproductive operator $\mathscr{R}_0$ and the net reproductive rate $\sigma(\mathscr{R}_0)$, we use system (1) without nonlinear term. In the worst case scenario for the system with dispersion, all migrants die before reaching next patch i.e. $D_{kj}(a) = 0$ for $a > 0$, $1 \leq k, j \leq N$, $k \neq j$. In this case we estimate $n_k(a,t)$ in the following way

$$\frac{\partial n_k(a,t)}{\partial t} + \frac{\partial n_k(a,t)}{\partial a} \geq -(\mu_k(a) + |D_{kk}(a)|)n_k(a,t)$$

and obtain

$$n_k(a,t) \geq \rho_k e^{-\int_0^a (\mu_k(v)+|D_{kk}(v)|)dv}\, da.$$

This implies that

$$\sigma(\mathscr{R}_0)\rho_k = \int_0^\infty m_k(a)n_k(a,t)\, da \geq \rho_k \int_0^\infty m_k(a)e^{-\int_0^a (\mu_k(v)+|D_{kk}(v)|)dv}\, da$$

and

$$\sigma(\mathscr{R}_0) \geq \max_{1\leq k\leq N} \int_0^\infty m_k(a)e^{-\int_0^a (\mu_k(v)+|D_{kk}(v)|)dv}\, da \tag{9}$$

On the other hand, the system of isolated patches corresponds to the problem (1)–(3) with $D_{kj}(a,t) = 0$ for $a, t > 0$ and $1 \leq k, j \leq N$. The net reproductive rate on each patch is

$$\sigma_k = \int_0^\infty m_k(a)e^{-\int_0^a \mu_k(v)dv}\, da.$$

If $\sigma_k \leq 1$ for some k, extinction of population on patch k is imminent. However, according to (9), it follows that $\sigma(\mathscr{R}_0) > 1$ if there exists at least one patch such that $\sigma_k > 1$ and $|D_{kk}(a)|$ is small enough. This means that persistence of population on all patches is possible under assumption that population would persist on at least one patch and that the number of individuals emigrating from this patch is sufficiently small.

5 Irregularly varying environment

Previously, we have established conditions that ensures persistence of population in time-independent case and for periodically changing environment. Here we establish asymptotic behavior of the solution to the main model in the general time-dependent case in two ways: by using periodic functions to formulate upper and lower bounds and by finding upper and lower boundaries for the sum of newborns on all patches and for the total population by

relaxing conditions for dispersion. In the first case, we will find conditions for extinction or persistence of population under assumption that the vital rates, regulating function and dispersion coefficients are bounded by periodic functions. In the second case we obtain a crude estimate of newborns density and total population which will not depend on migration pattern.

Suppose that the vital rates, regulating function and dispersion coefficients have periodic bounds, or, in other words, that the following estimates hold for the structure coefficients and large t:

$$\begin{aligned} m^-(a,t) \le m(a,t) \le m^+(a,t), \\ \mu^+(a,t) \le \mu(a,t) \le \mu^-(a,t), \text{ etc} \end{aligned} \tag{10}$$

where $m_k^\pm$, $\mu_k^\pm$, $L_k^\pm$ and $D_{kj}^\pm$ are some T-periodic functions. Instead of the original problem (1), we study two associated periodic problems with parameters defined by (10). It is natural to expect that $n(0,t)$ will be bounded from above and below by $\rho^\pm(t)$ for sufficiently large t, where $n(0,t)$ is a number of newborns in the original problem and $\rho^\pm(a,t)$ are solutions to the characteristic equations related to periodic problems

$$\widetilde{\mathcal{K}}^\pm \rho(t) = \rho(t), \quad t \in \mathbb{R},$$

where operators $\widetilde{\mathcal{K}}^\pm$ are defined componentwise by

$$(\widetilde{\mathcal{K}}^\pm \rho)_k(t) := \int_0^\infty m_k^\pm(a,t)\Phi_k^\pm(a,t-a;\rho)\,da, \quad t \in \mathbb{R}, \quad 1 \le k \le N. \tag{11}$$

The functions $\Phi^\pm(x,y;\rho)$ are unique solutions to the initial value problems

$$\begin{cases} \frac{d}{dx}h_k^\pm(x) &= -\mu_k^\pm(x,x+y)\left(1+\frac{h_k^\pm(x)}{L_k^\pm(x,x+y)}\right)h_k^\pm(x) + \sum_{j=1}^N D_{kj}^\pm(x,x+y)h_j^\pm(x), \\ h^\pm(0) &= \rho(y), \end{cases} \tag{12}$$

where the coefficients are T-periodic and satisfy condition (10) and the initial conditions are given by a vector-function $\rho \in C(\mathbb{R}_+, \mathbb{R}_+^N)$ such that $\rho(t+T) = \rho(t), \quad t \in \mathbb{R}$. Then

$$\rho^\pm(t) = \mathcal{K}^\pm \rho^\pm(t) + \mathcal{F}^\pm f(t),$$

where operators $\mathcal{K}^\pm$ and $\mathcal{F}^\pm$ are defined component-wise by

$$\begin{aligned} (\mathcal{K}^\pm \rho)_k(t) &= \int_0^t m_k^\pm(a,t)\Phi_k^\pm(a,t-a;\rho)\,da, \\ (\mathcal{F}^\pm f)_k(t) &= \int_t^\infty m_k^\pm(a,t)\Psi_k^\pm(a,a-t;f)\,da. \end{aligned}$$

The corresponding net reproductive operators and net reproductive rates are denoted by $\widetilde{\mathscr{R}}_0^\pm$ and $\sigma(\widetilde{\mathscr{R}}_0^\pm)$. Then the next result states that the number of newborns in irregularly changing environment can be bounded from above and below by the number of newborns in the associated periodically changing environments.

 ISBN: 978-84-608-6082-2

Theorem 4. *Let $\rho(t)$ be a solution to equation (4). If $\sigma(\widetilde{\mathscr{R}}_0^+) \leq 1$, then $\rho(t) \to 0$ as $t \to \infty$. If $\sigma(\widetilde{\mathscr{R}}_0^-) > 1$, then*

$$\rho^-(t) - \varepsilon \leq \rho(t) \leq \rho^+(t) + \varepsilon \quad \text{for large } t, \tag{13}$$

where $\rho^{\pm}(t)$ are solutions to (11) and ε is an arbitrary positive number.

References

[1] L. Allen, *Persistence and extinction in single-species reaction-diffusion models*, Bulletin of Mathematical Biology **45** (1983), 209–227.

[2] P. Amarasekare, *The role of density-dependent dispersal in source-sink dynamics*, Journal of Theoretical Biology **226** (2004), 159168.

[3] O. N. Bjrnstad and B. T. Grenfell, *Noisy clockwork: Time series analysis of population fluctuations in animals*, Science Translational Medicine **293 (5530)** (2001), 638–643.

[4] J. Chi and L. Chen, *The effect of diffusion on the time varying logistic population growth*, Computers Math. Applic. **36** (1998), 1–9.

[5] ———, *Permanence and extinction in logistic and lotka-volterra systems with diffusion*, Journal of Mathematical Analysis and Applications **258** (2001), 512–535.

[6] M. Chipot, *On the equations of age-dependent population dynamics*, Arch. Rational Mech. Anal. **82** (1983), no. 1, 13–25.

[7] ———, *A remark on the equation of age-dependent population dynamics*, Quarterly of Applied Mathematics **42** (1984), no. 2, 221–224.

[8] P. C. Dias, *Sources and sinks in population biology*, Trends in Ecology and Evolution **11** (1996), 326–330.

[9] C. A. Schmidt-Wellenburg et. al., *Trade-off between migration and reproduction: does a high workload affect body condition and reproductive state?*, Behavioral Ecology **doi:10.1093/beheco/arn066** (2008).

[10] P. A. Guerra, *Evaluating the life-history trade-off between dispersal capability and reproduction in wing dimorphic insects: a meta-analysis*, Biological Reviews **86** (2011), 813835.

[11] M. E. Gurtin and R. C. MacCamy, *Nonlinear age-dependent population dynamics*, Arch. Rat. Mech. Anal. **54** (1974), 281–300.

[12] A. Hastings, *Complex interactions between dispersal and dynamics: Lessons from coupled logistic equations*, Ecology **44** (1993), 1362–1372.

[13] P. Kareiva and U. Wennergren, *Connecting landscape patterns to ecosystem and population processes*, Nature **373** (1995), 299–302.

[14] M. A. Krasnosel′skiĭ and P. P. Zabreĭko, *Geometrical methods of nonlinear analysis*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 263, Springer-Verlag, Berlin, 1984, Translated from the Russian by Christian C. Fenske. MR 736839

[15] S. Mole and A. Zera, *Differential allocation of resources underlies the dispersal-reproduction trade-off in the wing-dimorphic cricket, gryllus rubens*, Oecologia **93** (1993), 121–127.

[16] J. W.-H. So, J. Wu, and X. Zou, *Structured population on two patches: modeling dispersal and delay*, J. Math. Biol. **43** (2001), 37–51.

[17] J. H. Steele, *A comparison of terrestrial and marine ecological systems*, Nature **313** (1985), 355–358.

[18] Y. Takeuchi, *Diffusion effect on stability of lotka-volterra models*, Bul. Math. Bio. **48** (1986), 585–601.

[19] _____, *Global stability in generalized lotka-volterra diffusion systems*, Journal of Math. Analysis and Appl. **116** (1986), 209–221.

[20] A. Terry, *Dynamics of structured population on two patches*, Journal of Mathematical Analysis and Applications **378** (2011), 1–15.

[21] P. Weng, C. Xiao, and X. Zou, *Rich dynamics in a non-local population model over three patches*, Nonlinear Dyn. **59** (2010), 161–172.

[22] V. Kozlov, S Radosavljevic, and U. Wennergren, *Large time behavior of logistic age-structured population model in changing environment*, submitted.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

m-adic Residue Codes over $F_q[v]/(v^2 - v)$ and DNA Codes

Ferhat Kuruz[1], **Elif Segah Oztas**[1] **and Irfan Siap**[1]

[1] *Department of Mathematics, Faculty of Art and Sciences, Yıldız Technical University*

emails: kuruz@yildiz.edu.tr, esoztas@yildiz.edu.tr, isiap@yildiz.edu.tr

Abstract

In this study we explore the structure of m-adic residue codes over $F_q[v]/(v^2 - v)$. We show that the generators of m-adic residue codes can be used to generate reversible DNA codes with a special automorphism and special sets over $F_{4^{2k}}[v]/(v^2 - v)$.

Key words: m-adic residue codes, polyadic codes, cyclic codes, DNA codes
MSC 2000: MSC 94B05, MSC 94B15

1 Introduction

Quadratic residue codes are an important class of cyclic codes. Some researchers has worked on further generalizations of these families of codes ([10], [9], [4], [14], [5]). Especially, m-adic residue codes are a generalization of this codes [6]. First, Pless and Brualdi has defined polyadic codes [4], later Pless has studied on polyadic codes via idempotent generators and specific ideals [10]. After these works, Job has defined m-adic residue codes in terms of generator polynomials over fields [6].

In [2] DNA computing studies is initiated by Leonard Adleman where a solution to and NP-complete problem is presented by employing DNA molecules.

DNA sequences consist of four bases (nucleotides) that are (A) Adenine, (G) Guanine, (T) Thymine and (C) Cytosine. DNA strands are formed by the famous Watson Crick complement (WCC) rule. "A" pairs with "T" and "G" pairs with "C" according to the rule WCC. We represent the WCC pairing as $A^c = T, T^c = A, G^c = C$ and $C^c = G$.

The error correction and detection quality observed in DNA strands mainly based on WCC property has attracted the attention on studying algebraic codes that enjoy these similar properties. Hence, studies on these directions have been one of the main focuses in algebraic coding theory. Such algebraic codes are referred to as DNA codes. To mention

some of these studies but surely not all, DNA codes are studied over $F_4([1])$,F_{16} ([11]), $F_{4^{2k}}$ ([12]), the chain rings $F_2[u]/(u^2-1)$ ([13]), $F_2[u]/(u^4-1)$ ([15]) and very recently over non-chain rings $F_4[v]/(v^2-v)$ ([3]). Due to the complex and still not well understood the structure of the DNA, researchers have restricted their studies to specific (local) regions of DNA strands such as protein, binding sites that has important role in the protein production process. In [8], reversible complement 8-bases (8-mers) are observed intensively in some specific and important regions of DNA.

In this work, we extend the definition of m-adic residue codes over a non-chain ring $(F_q/(v^2-v))$ in terms of idempotent generators. After exploring the structure of these codes, we introduce a special automorphism and a generator set to construct reversible DNA codes by means of the generators of m-adic codes over $F_{4^{2k}}/(v^2-v)$.

2 Preliminaries

Definitions in this section are compiled by help of [7] and [6]. Let q be a prime and F_q be a field with q elements. A subset C of F_q^n is called a code and a subspace of F_q^n is called a linear code. The number of non-zero coordinates of a codeword $x=(x_0,x_1,\ldots,x_{n-1})\in C$ is called Hamming weight of the codeword. Hamming distance between $x=(x_0,x_1,\ldots,x_{n-1})\in C$ and $y=(y_0,y_1,\ldots,y_{n-1})\in C$ is $d_H(x,y)=w_H(x-y)$ and minimum distance of C is $d_H(C-\{0\})=min\{d_H(x-y):x,y\in C, x\neq y\}$. A linear code of length n, dimension k and minimum distance d over F_q is referred to as an $[n,k,d]_q$ code. If further $(a_{n-1},a_0,\ldots,a_{n-2})\in C$ for all $(a_0,a_1,\ldots,a_{n-1})\in C$, C is called a cyclic code.
Let C_1 and C_2 be two cyclic codes. Let $g_1(x)$ and $g_2(x)$ be the generator polynomials C_1 and C_2 respectively and further $e_1(x)$ and $e_2(x)$ be idempotent generators of C_1 and C_2 respectively. Then, a generator polynomial of C_1+C_2 is $gcd(g_1(x),g_2(x))$ and an idempotent generator of C_1+C_2 is $e_1(x)+e_2(x)-e_1(x)e_2(x)$. Further, a generator polynomial of $C_1\cap C_2$ is $lcm(g_1(x),g_2(x))$ and an idempotent generator of $C_1\cap C_2$ is $e_1(x)e_2(x)$. Let $C_1\subseteq C_2$, then the complementary code of C_1 relative to C_2 is denoted by $\overline{C_1}$ having property that $C_1+\overline{C_1}=C_2$, and $C_1\cap\overline{C_1}=\overline{0}$. The complementary code of C_1 relatively to V (the whole space) is the complement of C_1. If $\sum_{i=0}^{n-1}v_i=0$ for $v=(v_0,v_1,\ldots,v_{n-1})\in V$, v is called even-like and otherwise odd-like. If all codewords of a code are even-like, then this code is an even-like code; otherwise an odd-like code. Let E be the set of all even-like vectors and $h(x)$ is the all one coefficient polynomial corresponding to $(1,1,\ldots,1)$. Then, the dimension of E is $n-1$ and this code a cyclic code with idempotent generator $1-(1/n)h(x)$.

Definition 1 *[6] Let p be a prime and b be a primitive element of Z_p^*. If $m\geq 2$, $m\in Z$ and $m|(p-1)$ then the set of nonzero m-adic residues modulo p is $Q_0=\{a^m : a\in Z_p\}$. Further, we let $Q_i=b^iQ_0$..*

Example 2 *Let $p=19$. 2 is a primitive element of Z_{19}^*. Since $6|19-1$, we can take $m=6$.*

Then, $Q_0 = \{1, 7, 11\}$, $Q_1 = \{2, 3, 14\}$, $Q_2 = \{4, 6, 9\}$, $Q_3 = \{8, 12, 18\}$, $Q_4 = \{5, 16, 17\}$, $Q_5 = \{10, 13, 15\}$.

Definition 3 *[6] Let p be a prime and q be prime power such that $gcd(p, q) = 1$. Let b be a primitive element of Z_p^* and γ be a primitive pth root of unity in some field extension of F_q. Let Q_0 be the set of nonzero m-adic residues modulo p and $Q_i = b^i Q_0$. If q is an m-adic residue modulo p, i.e. $q \in Q_0$, then the codes generated by polynomials $g_i(x) = \frac{x^p-1}{\prod_{k\in Q_i} x-\alpha^k}$ $(i = 0, 1, \ldots, m-1)$ are called even-like families of m-adic residue codes of class I of length p over F_q.*

The other families of m-adic residue codes defined below are derivations of even-like family of m-adic residue codes of class I.

Definition 4 *[6]*

- *The codes generated by polinomials $\widehat{g_i}(x) = \prod_{k\in Q_i} x - \alpha^k$ $(i = 0, 1, \ldots, m-1)$ are called odd-like class I m-adic residue codes of length p over F_q and the code generated by $\widehat{g_i}(x)$ is the complement of the code generated by $g_i(x)$.*
- *The codes generated by polinomials $h_i(x) = (x-1)\widehat{g_i}(x)$ $(i = 0, 1, \ldots, m-1)$ are called even-like class II m-adic residue codes of length p over F_q and the code generated by $h_i(x)$ is the complementary code of the code generated by $g_i(x)$ relative to E.*
- *The codes generated by polinomials $\widehat{h_i}(x) = \frac{g_i(x)}{x-1}$ $(i = 0, 1, \ldots, m-1)$ are called odd-like class II m-adic residue codes of length p over F_q and these codes are the complements of the codes generated by $h_i(x)$.*

Example 5 *Even-like class I 4-adic residue codes of length 17 over F_4 are*

$$C_0 = \langle g_0(x) = 1 + x + w^2x^2 + x^4 + w^2x^5 + wx^6 + wx^7 + w^2x^8 + x^9 + w^2x^{11} + x^{12} + x^{13}\rangle,$$

$$C_1 = \langle g_1(x) = 1 + wx + wx^2 + w^2x^3 + x^4 + wx^6 + wx^7 + x^9 + w^2x^{10} + wx^{11} + wx^{12} + x^{13}\rangle,$$

$$C_2 = \langle g_2(x) = 1 + x + wx^2 + x^4 + wx^5 + w^2x^6 + w^2x^7 + wx^8 + x^9 + wx^{11} + x^{12} + x^{13}\rangle,$$

and

$$C_3 = \langle g_3(x) = 1 + w^2x + w^2x^2 + wx^3 + x^4 + w^2x^6 + w^2x^7 + x^9 + wx^{10} + w^2x^{11} + w^2x^{12} + x^{13}\rangle$$

and minimum distance of these codes is 12.

Theorem 6 *[6] Let C be an arbitrary m-adic residue code with generating idempotent e. Then e is a linear combination of the polynomials $l_0(x)$, $l_1(x)$,$\ldots$,$l_{m-1}(x)$ and 1 over F_q where $l_i(x) = \sum_{k\in Q_i} x^k$.*

 ISBN: 978-84-608-6082-2

A polynomial is called palindromic if the order of its coefficients is reversed then it is still equal to itself. Later we are going to use the palindromic property of the generators of these families of codes. Here we first give a property of m-adic residue codes related to palindromic property.

Proposition 7 *If $q = 2^k$ where $k \geq 2 \in Z$, then the generator polynomials of the m-adic residue codes over F_q of length p are palindromic.*

In the following definition we give the definition of necessary notions.

Definition 8 *Let C be a code of length n over a finite alphabet. If $c^r = (c_{n-1}, c_{n-2}, \ldots, c_1, c_0) \in C$ for all $c = (c_0, c_1, \ldots, c_{n-1}) \in C$, then C is called a reversible code.*

3 m-adic residue codes over $F_q[v]/(v^2 - v)$

Pless ([10]) has defined idempotent generators of m-adic codes and she has given properties of these generators. Here, we extend these ideas of idempotent generators to m-adic residue codes over the ring $F_q[v]/(v^2 - v)$ and we identify idempotent generators for all classes of m-adic residue codes over this ring.

Proposition 9 *Let p be a prime and q be a prime power. Take a positive integer m such that $m|(p-1)$ and let a and q be m-adic residue modulo p, i.e $a, q \in Q_0$. If e_i and e_j are idempotents in $F(q)[x]/(x^p - 1)$, then $ve_i + (1-v)e_j$ are idempotents in $(F_q[v]/(v^2 - v))[x]/(x^p - 1)$.*

Proposition 10 *Let e_i and e_j be idempotent generators of an even-like class I m-adic residue code of length p over F_q such that $i, j \in \{0, 1, \ldots, m-1\}$. Let $E_k = ve_i + (1-v)e_j \in F_q[v]/(v^2 - v)[x]/(x^p - 1)$ and $h = 1 + x + x^2 + \ldots + x^{p-1}$. Then $\mu_a(E_k)$ is idempotent, $E_iE_j = 0$ $(i \neq j)$ and $E_0 + E_1 + \ldots + E_{m-1} = 1 - h$.*

Since idempotent elements $E_k = ve_i + (1-v)e_j$ satisfy these properties, we can take them as idempotent generators of m-adic residue codes over $F_q[v]/(v^2 - v)$.

Definition 11 *The codes generated by idempotent elements $E_k = ve_i + (1-v)e_j$ are called class I even-like m-adic residue codes of length p over $F_q[v]/(v^2 - v)$ where e_i and e_j are idempotent generators of class I even-like m-adic residue codes over F_q.*

Proposition 12 *Let E_i be the idempotent generator of a class I even-like m-adic residue codes of length p over $F_q[v]/(v^2 - v)$. Take $E'_i = 1 - E_i$ where $i, j \in \{0, 1, \ldots, m-1\}$. Then $\mu_a(E'_i) = E'_j$ $(i \neq j)$, $E'_i + E'_j - E'_iE'_j = 1$ $(i \neq j)$ and $E'_0E'_1 \ldots E'_{m-1} = h$.*

 ISBN: 978-84-608-6082-2

Definition 13 *The codes generated by idempotent element $E_i' = 1 - E_i$ are called class I odd-like m-adic residue codes of length p over $F_q[v]/(v^2 - v)$ where $i \in \{0, 1, \ldots, m-1\}$.*

Proposition 14 *Let E_i be the idempotent generator of a class I even-like m-adic residue codes of length p over $F_q[v]/(v^2 - v)$. Take $F_i = 1 - h - E_i$ where $i, j \in \{0, 1, \ldots, m-1\}$. Then $\mu_a(F_i) = F_j$ $(i \neq j)$, $F_i + F_j - F_iF_j = 1 - h$ $(i \neq j)$ and $F_0F_1 \ldots F_{m-1} = 0$.*

Definition 15 *The codes generated by idempotent element $F_i = 1 - h - E_i$ are called class II even-like m-adic residue codes of length p over $F_q[v]/(v^2 - v)$ where $i \in \{0, 1, \ldots, m-1\}$.*

Proposition 16 *Let E_i and F_i be class I and class II even-like m-adic residue code of length p respectively. Let $F_i' = 1 - F_i = h + E_i$ where $i, j \in \{0, 1, \ldots, m-1\}$. Then $\mu_a(F_i') = F_j'$ $(i \neq j)$, $F_i'F_j' = h$ $(i \neq j)$ and $F_0' + F_1' + \ldots + F_{m-1}' = 1 - (m-1)h$.*

Definition 17 *The codes generated by idempotent element $F_i' = 1 - F_i = h + E_i$ are called class II odd-like m-adic residue codes of length p over $F_q[v]/(v^2 - v)$ where $i \in \{0, 1, \ldots, m-1\}$.*

Example 18 *Idempotent generators of class I even-like 4-adic residue codes of length 17 over F_4 are $e_0 = l_0 + wl_1 + l_2 + w^2l_3$, $e_1 = wl_0 + l_1 + w^2l_2 + i_3$, $e_2 = l_0 + w^2l_1 + l_2 + wl_3$ and $e_3 = w^2l_0 + l_1 + wl_2 + l_3$ where $l_0 = x + x^4 + x^{13} + x^{16}$, $l_1 = x^3 + x^5 + x^{12} + x^{14}$, $l_2 = x^2 + x^8 + x^9 + x^{15}$, $l_3 = x^6 + x^7 + x^{10} + x^{11}$. If we take $E_0 = ve_0 + (1-v)e_2$, then idempotent generators of a class I even-like 4-adic residue codes of length 17 over $F_4[v]/(v^2 - v)$ are $E_0 = ve_0 + (1 - v)e_2$, $E_1 = \mu_3(E_0) = ve_3 + (1 - v)e_1$, $E_2 = \mu_3(E_1) = ve_2 + (1 - v)e_0$, $E_3 = \mu_3(E_2) = ve_1 + (1 - v)e_3$(also $E_0 = \mu_3(E_3)$).*
Let $g_i(x)$ be the generator polynomial of corresponding to E_i idempotent generator, then
$g_0(x) = 1 + x + (v + w)x^2 + x^4 + (v + w)x^5 + (v + w^2)x^6 + (v + w^2)x^7 + (v + w)x^8 + x^9 + (v + w)x^{11} + x^{12} + x^{13}$,
$g_1(x) = 1 + (v + w)x + (v + w)x^2 + (v + w^2)x^3 + x^4 + (v + w)x^6 + (v + w)x^7 + x^9 + (v + w^2)x^{10} + (v + w)x^{11} + (v + w)x^{12} + x^{13}$,
$g_2(x) = 1 + x + (v + w^2)x^2 + x^4 + (v + w^2)x^5 + (v + w)x^6 + (v + w)x^7 + (v + w^2)x^8 + x^9 + (v + w^2)x^{11} + x^{12} + x^{13}$,
$g_3(x) = 1 + (v + w^2)x + (v + w^2)x^2 + (v + w)x^3 + x^4 + (v + w^2)x^6 + (v + w^2)x^7 + x^9 + (v + w)x^{10} + (v + w^2)x^{11} + (v + w^2)x^{12} + x^{13}$.
This codes are $[17, 4, 8]$ codes.

Proposition 19 *Let p be a prime and q be a power of 2. Take an $m \in Z^+$ such that $m|(p-1)$ and let $a, q \in Q_0$. If e_i and e_j are idempotent generators of m-adic residue codes of length p over F_q, then $v(\sum_{S\subseteq, i\in S} e_i) + (1-v)(\sum_{P\subseteq I, j\in P} e_j)$'s are idempotents in the ring $(F_q[v]/(v^2 - v))[x]/(x^p - 1)$.*

Example 20 *If we choose* $E_0 = v(e_0 + e_1) + (1 - v)(e_1 + e_2)$ *as in the previous example, then* $E_1 = v(e_3 + e_0) + (1 - v)(e_0 + e_1)$, $E_2 = v(e_2 + e_3) + (1 - v)(e_3 + e_0)$ *and* $E_3 = v(e_1 + e_2) + (1 - v)(e_2 + e_3)$.
Let $g_i(x)$ *be the generator polynomial of corresponding to* E_i *idempotent generator, then*
$g_0(x) = 1+w^2x+(v+w)x^2+(vw+w)x^3+vwx^4+vwx^5+(vw+w)x^6+(v+w)x^7+w^2x^8+x^9$,
$g_1(x) = 1+(v+w^2)x+w^2x^2+vw^2x^3+(vw+w)x^4+(vw+w)x^5+vw^2x^6+w^2x^7+(v+w^2)x^8+x^9$,
$g_2(x) = 1 + wx + (v + w^2)x^2 + (vw^2 + w^2)x^3 + vw^2x^4 + vw^2x^5 + (vw^2 + w^2)x^6 + (v + w^2)x^7 + wx^8 + x^9$,
and $g_3(x) = 1 + (v + w)x + wx^2 + vwx^3 + (vw^2 + w^2)x^4 + (vw^2 + w^2)x^5 + vwx^6 + wx^7 + (v + w)x^8 + x^9$.
These codes are all $[17, 8, 8]$ *codes.*

4 Reversible DNA codes over $F_{4^{2k}}/(v^2 - v)$

In this section we use the m-adic residue codes and a general form of ψ-set with an automorphism introduced by the authors to solve reversibility problem for DNA codes over $R_{2k} = F_{4^{2k}}/(v^2 - v)$. The ψ-set originally is introduced in [3] over $F_4/(v^2 - v)$.

In order to explain the reversibility problem we give an example. Let (a_1, a_2, a_3) be a codeword corresponding to ATGGCTGATGAG (a 12-string) where a_1 $\rightarrow$ATGG, a_2 $\rightarrow$CTGA, α_3 $\rightarrow$TGAG and $a_1, a_2, a_3 \in R_4$. The reverse of (a_1, a_2, a_3) is (a_3, a_2, a_1), and (a_3, a_2, a_1) corresponds to TGAGCTGAATGG. However, TGAGCTGAATGG is not the reverse of ATGGCTGATGAG. Indeed, the reverse of ATGGCTGATGAG is GAGTAGTCGGTA. So this concrete example reveals the fact that the reverse in the ring form of the codewords does not lead to the reverse of the DNA strings.

R is a commutative non-chain ring where $v^2 = v$. By Chinese Remainder Theorem we can decompose R as follows: $R = vF_{4^{2k}} \oplus (1 - v)F_{4^{2k}}$. We define a Gray map:

$$\begin{aligned} \phi : R &\rightarrow F_{4^{2k}}^2 \\ a + vb &\rightarrow (a + b, a). \end{aligned} \tag{1}$$

θ which used to convert the elements of the $F_{4^{2k}}$ to DNA strings of length $2k$ as in the Tables defined in [11, 12]. Especially, the DNA table for F_{16} that matches the field elements with DNA doubles is presented in [11]. More general θ_1 is used to convert the elements of the R_{2k} to DNA strings of length $4k$. Let $a+vb \in R_{2k}$. $\theta_1(a+vb) = (\theta(a+b), \theta(a))$. Θ is used to convert the codeword to DNA. Let $c = (c_0, c_1, ..., c_{n-1})$ and $\Theta(c) = (\theta_1(c_0), \theta_1(c_1), ..., \theta_1(c_{n-1}))$.

Example 21 *Let* $\beta = \alpha^3 + \alpha^6 v \in R_2$ *and* $\phi(\beta) = (\alpha^2, \alpha^3)$ *then* $\theta_1(\beta) = (\theta(\alpha^2), \theta(\alpha^2)) = GCAG$

Example 22 *Let* $c = (\alpha^3 + \alpha^6 v, \alpha^3 + \alpha^9 v)$ *be a codeword of a code.* $\phi(\alpha^3 + \alpha^6 v) = (\alpha^2, \alpha^3)$ *and* $\phi(\alpha^3+\alpha^9 v) = (\alpha, \alpha^3)$. $\Theta(c) = (\theta_1(\alpha^3+\alpha^6 v), \theta_1(\alpha^3+\alpha^9 v)) = (\theta(\alpha^2), \theta(\alpha^3), \theta(\alpha), \theta(\alpha^3)) = GCAGATAG$

Here, we introduce a new automorphism over R that leads to obtaining the DNA reverses of elements in R. This is also a generalization of the map introduced θ in [11, 12].

$$\begin{aligned} \psi : R &\to R \\ a + vb &\to a^{4^k} + (1+v)b^{4^k} \\ &= (a+b)^{4^k} + vb^{4^k}. \end{aligned} \tag{2}$$

Example 23 *Let* $\beta = \alpha^3 + \alpha^6 v \in R_2$ *and* $\theta_1(\beta) = GCAG$ *then* $\psi(\beta) = \alpha^1 2 + \alpha^9(v-1) = \alpha^8 + v\alpha^9$ *and* $\theta_1(\psi(\beta)) = \theta_1(\alpha^8 + v\alpha^9) = (\theta(\alpha^{12}), \theta(\alpha^8)) = GACG$.

Definition 24 *Let* $g(x)$ *be a polynomial with* $\deg g(x) = t$ *over* R. *Let*

$$\Lambda_g = \{\Lambda_0, \Lambda_1, ..., \Lambda_{t-1}\}$$

where

$$\Lambda_i = \begin{cases} x^i g(x) & \text{if } i \text{ is even} \\ x^i \psi(g(x)) & \text{if } i \text{ is odd.} \end{cases}$$

The set Λ_g *is called the* ψ*-set.*

Theorem 25 *Let* $g(x)$ *be an idempotent generator of an* m*-adic residue codes over* $F_{4^{2k}}/(v^2 - v)$ *where* $\deg(x)$ *is odd. If* C *is a linear code generated by a* ψ*-set, then* $\Theta(C)$ *is a reversible DNA code.*

5 Conclusion

Here we construct m-adic residue codes over the non-chain $F_q[v]/(v^2 - v)$ and we explore the structure of DNA codes with ψ set of the generators of m-adic residue codes over $F_{4^{2k}}[v]/(v^2 - v)$. We further relate these findings to DNA codes.

References

[1] T. Abualrub, A. Ghrayeb and X. N. Zeng, *Consruction of cyclic codes over* F_4 *for DNA computing*, Journal ofthe Franklin Ins., **343** (2006), 448–457.

[2] L. Adleman, *Molecular computation of solutions to combinatorial problems*, Science **266** , (1994), 1021-1024.

[3] A. Bayram, E.S. Oztas and I. Siap, *Codes over* $F_4 + vF_4$ *and some DNA applications*, Designs, Codes and Cryptography, (2015), doi: 10.1007/s10623-015-0100-8.

[4] R. Brualdi, V. Pless, *Polyadic codes*, Discrete Applied Mathematics, Elsevier, **25** (1989) 3–71.

[5] X. Dong, L. Wenjie, Z. Yan, *Generating idempotents of cubic and quartic residue codes over field F_2*, Designs, Computer Engineering and Applications, North China Computing Technology Institute, **49** (2013) 41–44.

[6] V. R. Job, *m-adic residue codes*, Information Theory, IEEE Transactions on **38** (1992) 496–501.

[7] S. Ling and C. Xing, *Coding theory: a first course*, Cambridge University Press, (2004).

[8] J. Lichtenberg, A. Yilmaz, J. Welch, K. Kurz, X. Liang, F. Drews, K. Ecker, S. Lee, M. Geisler, E. Grotewold and L. Welch, *The word landscape of the non-coding segments of the Arabidopsis thaliana genome*, BMC Genomics, **10**, (2009), 463.

[9] F. MacWilliams, *Generalized quadratic residue codes*, Information Theory, IEEE Transactions on **24** (1978) 730–737.

[10] V. Pless, *Polyadic Codes*, Algebraic Combinatorial Theory (1988) 107–115.

[11] E. S. Oztas and I. Siap, *Lifted polynomials over F_{16} and their applications to DNA Codes*, Filomat, **27** (2013), 459–466.

[12] E.S. Oztas, I. Siap, *On a generalization of lifted polynomials over finite fields and their applications to DNA codes*, Int. J. Comput. Math., **92,** No. 9, (2015), 1976–1988.

[13] I. Siap, T. Abualrub and A. Ghrayeb, *Cyclic DNA codes over the ring $F_2[u]/(u^2-1)$ based on the deletion distance*, Journal of the Franklin Ins., **346** (2006) 731–740.

[14] A. J. van Zanten, A. Bojilov, S. M. Dodunekov, *Generalized residue and t-residue codes and their idempotent generators*, Designs, Codes and Cryptography, Springer, **75** (2015) 315–334.

[15] B. Yildiz, I. Siap, *Cyclic codes over $F_2[u]/(u^4-1)$ and applications to DNA codes*, Computers and Mathematics with Application, **63**, (2012) 1169 -1176.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Dynamics of a bouncing ball

Marek Lampart[1] **and Jaroslav Zapoměl**[2,3]

[1] *Department of Applied Mathematics & IT4Innovations,*
VŠB - Technical University of Ostrava, Czech Republic.

[2] *Department of Mechanics,*
VŠB - Technical University of Ostrava, Czech Republic.

[3] *Department of Dynamics and Vibrations,*
Institute of Thermomechanics, Czech Republic.

emails: marek.lampart@vsb.cz, jaroslav.zapomel@vsb.cz

Abstract

The main aim of the paper is to research dynamic properties of a mechanical system consisting of a ball jumping between a movable baseplate and a fixed upper stop. The model is constructed with one degree of freedom in the mechanical oscillating part. The ball movement is generated by the gravity force and harmonic oscillation of the baseplate in the vertical direction. The impact forces acting between the ball and plate and the stop are described by the nonlinear Hertz contact law. The ball motion is then governed by a set of two nonlinear ordinary differential equations. To perform their solving the Runge-Kutta method of the 4th order with adaptable time step was applied. As the main result it is shown that the systems exhibits regular, irregular and chaotic pattern for different choices of parameters using standard methods.

Key words: mechanical model, chaos tests, bifurcation, vibration
MSC 2000: 34H20, 34H10, 37N30

1 Introduction

A simple impact process model, that has big practical importance, has been extensively studied by physicists in past decades. In this model a small ball bounces vertically on a massively vibrating baseplate that produce a sequence of impacts arriving in the ball regular and irregular oscillatory motion.

This model has been researched in simulation way by [6] for impulsive noise, especially from metal-to-metal collisions, in influencing the noise levels produced in manufacturing operations. Later on, in [3] the same situation was investigated in experimental way. In [2] this model was simulated showing that for sufficiently large excitation velocities and a coefficient of restitution close to one, this deterministic system exhibits large families of irregular non-periodic solutions in addition to the expected harmonic and subharmonic motions.

This paper was motivated by [3–5] where electromechanical systems damped by impact element was introduced, for a purpose of attenuation of body vibrations. These vibrations are showing periodic, quasi periodic and chaotic patterns. The main problem is to detect which of those movements are regular or irregular, standard tools were used for the movement analysis and they gave unsatisfactory results. The carried out research extends the problem of impact dampers that were subjects of investigations in [3–5].

In this paper, a mechanical system formed by a ball bouncing between two stops is analyzed.

The investigated system consists of a vibrating baseplate (body 2, Fig. 1), a fixed upper stop (body 3 , Fig. 1), and of a bouncing ball (body 1, Fig. 1). The ball and the baseplate can move only in the vertical direction. The Hertz theory has been adopted to describe the contact stiffness between the ball and the stops.

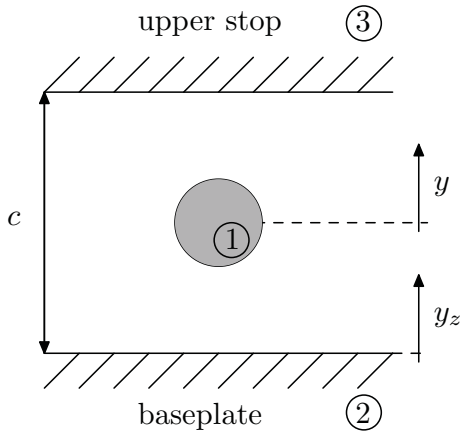


Figure 1: Model of vibrating system

The task was to analyze the influence of the systems parameters, especially the excitation frequency of the base plate, on a character of the ball movement. The investigated system has one degree of freedom. Its instantaneous position is defined by one generalized coordinate:

y - vertical position of the bouncing ball.

The equation of movement, that can be derived directly, has the following form:

$$m\ddot{y} = -mg + F_h + F_s \tag{1}$$

where F_h and F_s are the impact forces. They are defined as follows:

$$F_h = \begin{cases} -k_c(y-c) - b_c\dot{y} & \text{if } y_z > 0, \\ 0 & \text{if } F_h > 0, \end{cases} \tag{2}$$

$$F_s = \begin{cases} -k_c(y_z - y) - b_c(\dot{y}_z - \dot{y}) & \text{if } y < y_z, \\ 0 & \text{if } F_s < 0, \end{cases} \tag{3}$$

here ($\dot{}$) and ($\ddot{}$) denote the first and second derivative with respect to time, respectively.

The movement of the baseplate is described by

$$y_z(t) = A\,(1 - \mathrm{e}^{-\alpha t})\sin(\omega t) \tag{4}$$

as a rheonomic constraint, where A is the ground vibration amplitude, α is the constant determining how fast the vibration of the baseplate becomes a steady state and ω stands for the excitation frequency. All system parameters are summarized in Table 1.

It can be assumed, without loss of generality, that the ball is placed in the middle of the upper stop and the baseplate, taking the rest position. That is:

$$y(0) = 0.5\,c, \tag{5}$$

$$\dot{y}(0) = 0. \tag{6}$$

2 Main results

Main results were reached by numerical simulations of (1) for system parameters summarized in Table 1 where the excitation frequency ω was changed form 70 rad s^{-1} to 140 rad s^{-1}.

The character of movement of the ball is analyzed in detail. The first observation is shown in bifurcation diagram Fig. 2, where periodic as well as chaotic movements are visible for suitable choice of ω. In this figure it is also detected that for the excitation frequency $\omega \in [70, 108]$ there are no contacts of the ball with the upper stop, and for $\omega \in [109, 140]$ collisions of the ball with the upper stop appear.

It is visible form phase portraits (see Figs. 3 and 4) that the movement is periodic (e.g. T3 for $\omega = 88$ rad s^{-1}) and also chaotic (e.g. for $\omega = 130$ rad s^{-1}) and its character is changing with increasing excitation frequency.

The output paremeter of the 0-1 test for chaos can acquire only one of the values 0 or 1 which correspond to the regular and chaotic motions, respectively. More details can be

 ISBN: 978-84-608-6082-2

Table 1: Parameters of the system (1).

value	quantity	format	description
m	20	kg	mass of the bouncing ball
c	0.2	m	clearance between the baseplate and the upper stop
α	1	s^{-1}	parameter of the baseplate excitation
ω		rad s^{-1}	baseplate excitation frequency
A	1	mm	ground vibration amplitude
k_c	1×10^7	N m^{-1}	contact stiffness
b_c	100	N s m^{-1}	coefficient of contact damping

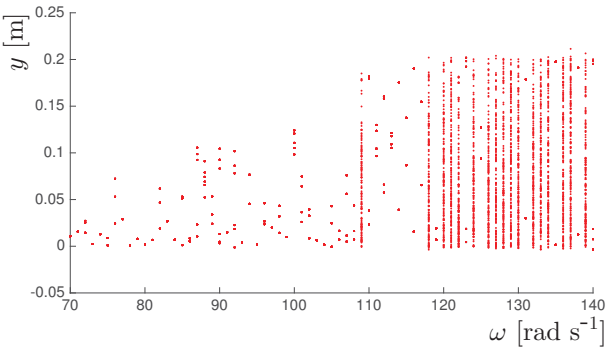


Figure 2: Bifurcation diafram of y in dependence on ω.

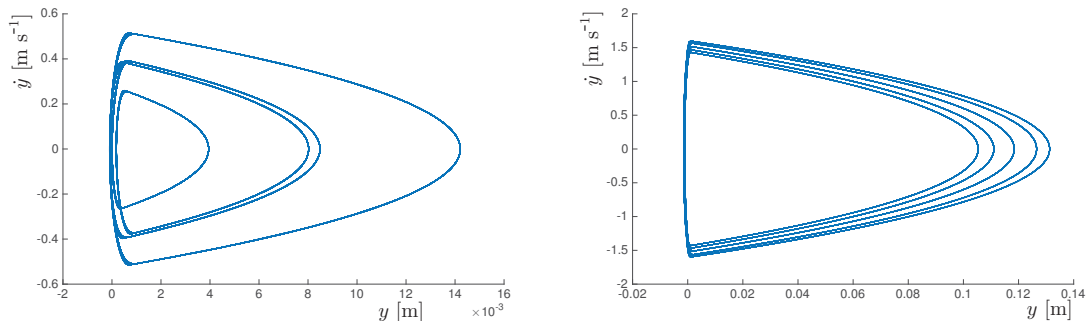


Figure 3: Phase portraits y versus $\dot{y}$ for $\omega = 75$ rad s^{-1} (left) and $\omega = 100$ rad s^{-1} (right).

 ISBN: 978-84-608-6082-2

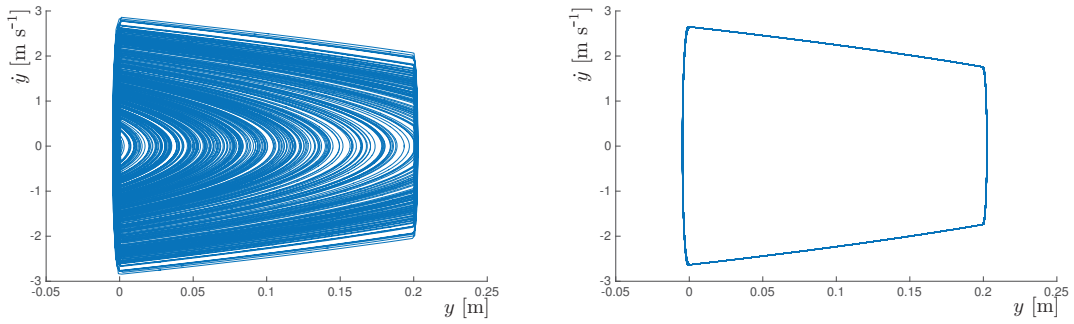


Figure 4: Phase portraits y versus $\dot{y}$ for $\omega = 130$ rad s^{-1} (left) and $\omega = 131$ rad s^{-1} (right).

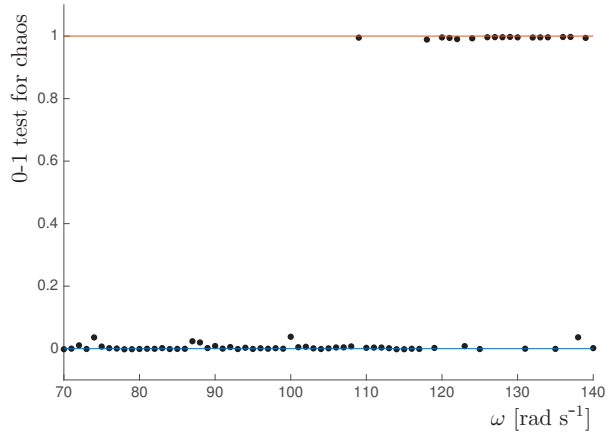


Figure 5: Output of the 0-1 test for chaos of y in dependence on ω.

found in [1]. The results of the 0-1 test for the range of investigated frequencies are shown in Fig. 6 where chaotic and non-chaotic movements were detected. Note, that the result of 0-1 test coincides with bifurcation diagram in Fig. 2 and with phase diagrams, for special choices of the excitation frequencies shown in Figs. 3 and 4).

3 Conclusions

In this paper, a mechanical systems with one degree of freedom has been investigated and the movement of the bouncing ball was analyzed. This model was inspired by a real problem of characterization of movements of an impact element of the impact damper that is used for the vibration attenuation of the electromechanical system.

The equation of movement was solved numerically using Runge-Kutta method implemented as *ode45* solver in Matlab.

It was observed that the movement is showing regular and also chaotic patterns for suitable choice of parameters, mainly excitation frequency of the baseplate played a key role here. For this purpose 0-1 test for chaos and bifurcation diagrams were used. The bouncing ball forced by sinusoidally vibrating baseplate was also getting collisions with the upper stop and this fact was also compared with the regularity of movement.

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602". The work was also supported by the Czech Science Foundation, Grant No. 15-06621S.

References

[1] G. A Gottwald, I. Melbourne, *A new test for chaos in deterministic systems*, Proc. R. Soc. London A **460** (2004) 603–611.

[2] J. P. Holmes, *The dynamics of repeated impacts with a sinusoidally vibrating table*, Journal of Sound and Vibration **84** (1982) 173–189.

[3] M. Lampart, J. Zapoměl, *Dynamics of the electromechanical system with impact element*, Journal of Sound and Vibration **332** (2013) 701–713.

[4] M. Lampart, J. Zapoměl, *Dynamic properties of the electromechanical system damped by an impact element with soft stops*, International Journal of Applied Mechanics **6** (2014) 1450016 (17 pages).

[5] M. Lampart, J. Zapoměl, *Vibration attenuation of an electromechanical system coupled with plate springs damped by an impact element*, International Journal of Applied Mechanics **7** (2015) 1550043 (14 pages).

[6] L. A. Wood, K. P. Byrne, *Analysis of a random repeated impact process*, Journal of Sound and Vibration **78** (1981) 329–345.

[7] L. A. Wood, K. P. Byrne, *Experimental investigation of a random repsated impact process*, Journal of Sound and Vibration **85** (1982) 53–69.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Modelling parts of branched skins using rational envelope surfaces

Miroslav Lávička[1,2] and Michal Bizzarri[1]

[1] *NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

[2] *Department of mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

emails: lavicka@kma.zcu.cz, bizzarri@ntis.zcu.cz

Abstract

Due to its technical importance, the operation of skinning has attracted the geometric modelling community in recent years. Especially spheres belong among the most used input shapes as their skins play an important role in various applications, e.g. computational chemistry, molecular biology, computer animation, and modelling of tubular surfaces. When branched skins of systems of spheres are constructed then the envelopes of suitable two-parametric systems of spheres must be considered. In this paper we solve this problem using the so-called rational envelope surfaces. The functionality of the designed algorithm is presented on some examples.

Key words: Skinning; rational surfaces; square-root parameterizations; skeletal structure; computational chemistry and biology

1 Introduction

The problem of *skinning* appears in various situations. Probably the best known is the application in computer animation – given a skeletal pose, then skinning algorithms are responsible for deforming the geometric skin to respond to the motion of the underlying skeleton. Skinning is a construction of a G^k (C^k) continuous interpolation curve/surface of an ordered sequence of planar or spatial shapes. This operation can be viewed as a particular analogy to the well-known interpolation of point data sets, which is one of the most crucial techniques in Computer Aided Geometric Design, cf. [6] and references therein.

Due to its technical importance, skinning has attracted the geometric modelling community in recent years and one can find several papers on this topic, see e.g. [15, 9, 1].

In this paper we will focus on a special case of ball skinning. The problem of ball skinning appears frequently for instance in the area of computational chemistry or molecular biology when surface meshes for molecular models are supposed to be generated. One can find several algorithms to skin a molecular model to produce a piecewise smooth surface. When dealing with a continuous family of balls, the skin is the envelope of the infinite set of the circles of intersection of two infinitely close spheres. In a discrete sense, skinning can be regarded as a part of the problem of computing envelopes of families of circles/spheres using the cyclographic mapping [12, 14]. Skinning is also closely related to representing shapes with the help of the associated medial axis/surface transforms [4, 11] and the theory of canal and pipe surfaces [10, 5].

We recall that from practical reasons (application of NURBS description which is nowadays a standard in computer aided geometric design) it is suitable to formulate such algorithms which produce piecewise rational skin surfaces. This is especially hard problem when branched skins shall be constructed, cf. [1]. In this paper we will present a simple method how to compute a special elements (parts) of rational branched skins using the recently introduced rational envelope shapes, i.e., medial axis/surface transforms allowing square roots in the radius function but guaranteeing the rationality of the associated envelopes, see [3].

2 Preliminaries

Firstly, we shortly recall some fundamental facts about skinning of balls/spheres and about rational envelope surfaces.

2.1 Skinning balls in 3D space

Following the approach from [9], we consider a given (and admissible) sequence of spheres $\Sigma = \{S_1, S_2, \ldots, S_n\}$. Our goal is to describe a G^1 spline surface $\mathcal{S}(\Sigma)$ skinning this system. As in [1], we will admit not only linear sequences of input spheres but also more complicated situations. Especially, we focus on configurations when branched skins shall be constructed.

When constructing the skin of spheres in the linear (sub)configuration then $\mathcal{S}(\Sigma)$ consists of the following elements: (i) parts of S_i obtained as the differences of S_i and the spherical caps determined by the contact circles; (ii) surfaces smoothly joining two consecutive spheres S_{i-1} and S_i along prescribed contact circles. For constructing the blending shapes (ii) we can use e.g. the algorithm formulated in [2]. It is based on using rational contour curves of canal surfaces for computing their rational parameterizations. It is beyond the scope of this paper to go into details and we refer e.g. to [1].

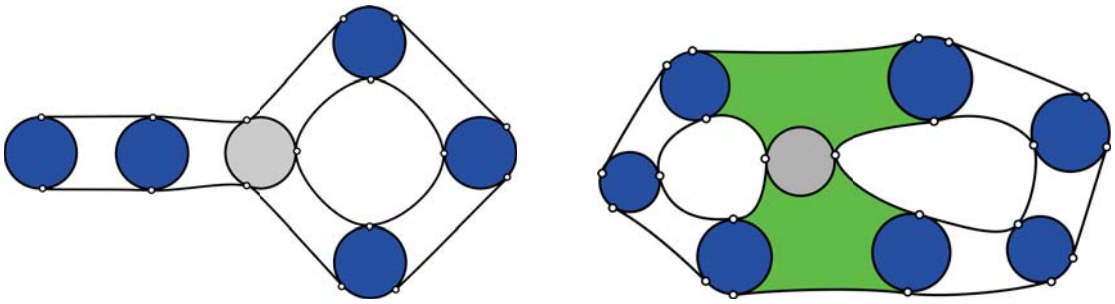


Figure 1: Branching of a skinning surface: on a particular sphere (left), or when a blend (green) between more than two spheres is needed (right). The gray sphere is connected via the skin with 3 (left), or 4 (right) neighboring spheres.

Nonetheless, the problem is more complicated when branching of skins is allowed. We recall that branching in this case means that there exists a sphere which is connected via the skin with more than two neighboring spheres, see Fig. 1 (the gray sphere). Then two types of situations can arise. Either, branching occurs on a particular sphere, Fig. 1 (left). In this case, the skin consists again of the elements of type (i) and (ii) only. Or, a blend between more than two spheres is needed, i.e., new type (iii) is needed – see the green shapes in Fig. 1 (right). Hence the rational envelopes of suitable two-parameter systems of spheres will be investigated in this paper in more detail. This problem was partially mentioned also in [1] however non-rational envelopes were applied.

2.2 Rational envelope surface in $\mathbb{R}^{3,1}$

When constructing rational branched skins then it is necessary to construct rational envelopes of suitable two-parametric systems of spheres. We will model this elements using RE surfaces.

Consider a spatial domain $\Omega \subset \mathbb{R}^3$ and the family of all inscribed spheres partially ordered with respect to inclusion of the associated balls. Then the *medial surface* (MS) of Ω is the locus of all centers $\mathbf{y}$ of maximal inscribed spheres and the *medial surface transform* (MST) of Ω is obtained by appending the corresponding sphere radius r to the medial surface, i.e., it consists of points $\overline{\mathbf{y}} = (\mathbf{y}, r) \in \mathbb{R}^{3,1}$.

For a C^1 segment $\overline{\mathbf{y}}(u,v) = (\mathbf{y}(u,v), r(u,v)) \subset \mathbb{R}^{3,1}$ of an MST, the corresponding boundary of the domain Ω is given by the envelope formula

$$\mathbf{x}^{\pm} = \mathbf{y} - r\,\mathbf{n}^{\pm}, \quad \mathbf{n}^{\pm} = \frac{(r_u G - r_v F)\,\mathbf{y}_u + (r_v E - r_u F)\,\mathbf{y}_v \pm (\mathbf{y}_u \times \mathbf{y}_v)\sqrt{\overline{EG} - \overline{F}^2}}{EG - F^2}, \quad (1)$$

where x_u denotes the partial derivatives of x w.r.t. the variable u, etc. The components $\overline{E}, \overline{F}, \overline{G}$ of the first fundamental form of $\overline{\mathbf{y}}(u,v)$ are computed using the indefinite *Minkowski inner product*

$$\langle \mathbf{u}, \mathbf{v} \rangle = u_1 v_1 + u_2 v_2 + u_3 v_3 - u_4 v_4, \tag{2}$$

whereas the components E, F, G of the first fundamental form of $\mathbf{y}(u,v)$ are determined using the standard Euclidean inner product in $\mathbb{R}^3$.

The so-called *MOS surfaces* (i.e., *Medial surfaces Obeying a certain Sum of squares condition*, see [7]) are characterized by the condition

$$\overline{EG} - \overline{F}^2 = \sigma^2(u,v), \tag{3}$$

where $\sigma(u,v) \in \mathbb{R}(u,v)$. This ensures that the envelope $\mathbf{x}^{\pm}$ is rational. Consequently, $\mathbf{x}^{\pm}$ possesses a normal vector field $\mathbf{n}^{\pm} = (\mathbf{x}^{\pm} - \mathbf{y})/r$ rationally parameterizing the unit sphere, i.e., $\mathbf{x}^{\pm}$ are rational surfaces with *Pythagorean normals* (*PN surfaces* for short), see [13]. Additionally, any rational MOS surface $\overline{\mathbf{y}}$ in $\mathbb{R}^{3,1}$ can be constructed starting from an (associated) rational PN surface $\mathbf{x}$ in $\mathbb{R}^3$ and a rational function r in the form

$$\overline{\mathbf{y}}(u,v) = \left(\mathbf{x} + r \frac{\mathbf{x}_u \times \mathbf{x}_v}{\sigma}, r \right). \tag{4}$$

We remark that $\mathbf{x}$ will play the role of $\mathbf{x}^+$ in what follows. However, the main problem is that the algorithms for interpolations with PN surfaces are relatively complicated (as they are often based on the dual approach, or on the reparameterizations).

Nonetheless, MOS surfaces are not the only MSTs yielding rational envelopes. Turning back to (1), we only have to guarantee that $r\mathbf{n}^{\pm}$ is rational. This brings us to a broader class of (generally non-rational) *RE surfaces*, i.e., *surfaces yielding Rational Envelopes*. Accordingly, we set $r(u,v)$ as the square root of some non-negative function $R(u,v)$. This leads to

$$r r_u = \frac{1}{2} R_u \in \mathbb{R}(u,v), \quad r r_v = \frac{1}{2} R_v \in \mathbb{R}(u,v). \tag{5}$$

Then the rationality of $r\mathbf{n}^{\pm}$ (and thus also of the envelope $\mathbf{x}^{\pm}$), cf. (1), is guaranteed by the condition

$$R(\overline{EG} - \overline{F}^2) = \sigma^2(u,v). \tag{6}$$

Additionally, any RE surface $\overline{\mathbf{y}}$ in $\mathbb{R}^{3,1}$ can be constructed starting from an (associated) rational surface $\mathbf{x}$ in $\mathbb{R}^3$ and a rational function f in the form

$$\overline{\mathbf{y}}(u,v) = \left(\mathbf{x} + f\,(\mathbf{x}_u \times \mathbf{x}_v), f\,|\mathbf{x}_u \times \mathbf{x}_v| \right), \tag{7}$$

In contrast to MOS surfaces it is easy to generate RE surfaces in the form $\overline{\mathbf{y}} = (\mathbf{y}, r = \sqrt{R})$.

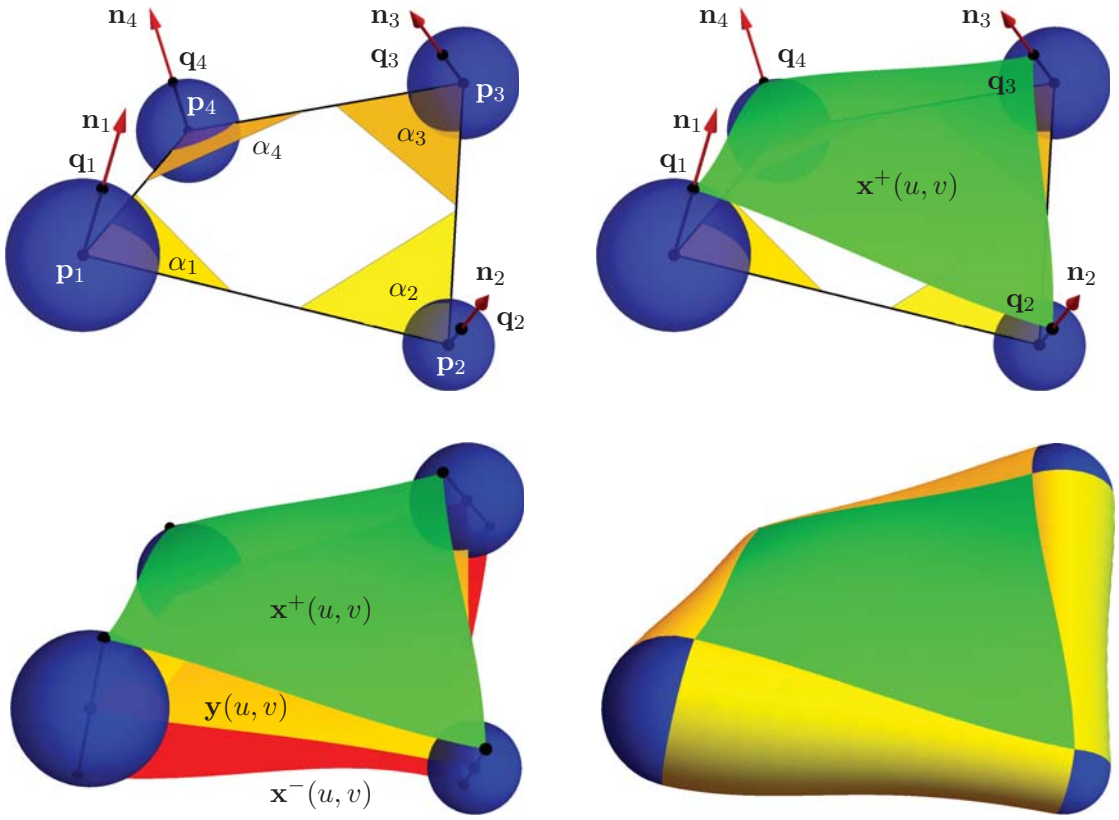


Figure 2: Construction of a rational skinning surface of four given spheres.

3 Skinning balls using rational envelope surfaces

Consider a system of n spheres in $\mathbb{R}^3$ (typically $n = 3, 4$) and our task is to find a rational skinning surface enveloping these spheres. This shape can be then used for constructing skins branched at this element. A main idea of our approach is to construct an RE surface as MST in $\mathbb{R}^{3,1}$ such that the corresponding envelope surface in $\mathbb{R}^3$ will give the wanted skin.

In particular, we start with constructing a suitable polynomial surface $\mathbf{x}(u, v) \subset \mathbb{R}^3$ considered as one branch (e.g. $\mathbf{x}^+(u, v)$) of the envelope surface. Next using a suitably chosen function $f(u, v)$ we obtain a polynomial MST $\overline{\mathbf{y}}(u, v) \subset \mathbb{R}^{3,1}$, cf. (7). Of course the second branch $\mathbf{x}^-(u, v)$ of the envelope surface associated to $\overline{\mathbf{y}}(u, v)$ will possess the rational description as well.

Let the spheres $\mathcal{S}_i$, $i = 1, ..., n$, be given as the points $\overline{\mathbf{p}}_i = (\mathbf{p}_i, r_i) \in \mathbb{R}^{3,1}$ such that for

each point $\overline{\mathbf{p}}_i$ the two points $\overline{\mathbf{p}}_{i-1}, \overline{\mathbf{p}}_{i+1}$ are considered as the associated neighboring points. For the sake of correctness we consider $\mathbf{p}_{-1} = \mathbf{p}_n$. First, we choose a tangent plane $\overline{\alpha}_i$ at each point $\overline{\mathbf{p}}_i$ by the tangent vectors

$$\overline{\mathbf{t}}_{i1} = \lambda_{i1}\left(\overline{\mathbf{p}}_{i-1} - \overline{\mathbf{p}}_i\right) = (\mathbf{t}_{i1}, \tau_{i1}), \quad \overline{\mathbf{t}}_{i2} = \lambda_{i2}\left(\overline{\mathbf{p}}_{i+1} - \overline{\mathbf{p}}_i\right) = (\mathbf{t}_{i2}, \tau_{i2}), \quad \lambda_{i1}, \lambda_{i2} \in \mathbb{R}. \tag{8}$$

Then α_i denotes the plane in $\mathbb{R}^3$ given by $\mathbf{p}_i, \mathbf{t}_{i1}, \mathbf{t}_{i2}$. Next using envelope formula (1) we obtain the associated end points $\mathbf{q}_i$ on the corresponding envelope

$$\mathbf{q}_i = \mathbf{p}_i - r_i \frac{(\tau_{i1}G_i - \tau_{i2}F_i)\,\mathbf{t}_{i1} + (\tau_{i2}E_i - \tau_{i1}F_i)\,\mathbf{t}_{i2} \pm (\mathbf{t}_{i1} \times \mathbf{t}_{i2})\sqrt{\overline{E}_i\overline{G}_i - \overline{F}_i^2}}{E_iG_i - {F_i}^2}, \tag{9}$$

where $E_i = |\mathbf{t}_{i1}|^2$, $F_i = |\mathbf{t}_{i2}|^2$, $G_i = \mathbf{t}_{i1} \cdot \mathbf{t}_{i2}$ and $\overline{E}_i = |\mathbf{t}_{i1}|^2 - \tau_{i1}^2$, $\overline{F}_i = |\mathbf{t}_{i2}|^2 - \tau_{i2}^2$, $\overline{G}_i = \mathbf{t}_{i1} \cdot \mathbf{t}_{i2} - \tau_{i1}\tau_{i2}$. Naturally, the normal vectors $\mathbf{n}_i$ of the envelope surface at $\mathbf{q}_i$ are

$$\mathbf{n}_i = \lambda_i(\mathbf{q}_i - \mathbf{p}_i), \qquad \lambda_i \in \mathbb{R}, \tag{10}$$

see Fig. 2 (top, left). We can construct a polynomial (triangular or quadrilateral) patch $\mathbf{x}(u, v)$ interpolating the points $\mathbf{q}_i$ and the associated normal vectors $\mathbf{n}_i$, see Fig. 2 (top, right). Next, using formula (7) we lift $\mathbf{x}(u, v)$ to $\mathbb{R}^{3,1}$ and as a result we arrive at the MST $\overline{\mathbf{y}}(u, v)$. Conditions for $\overline{\mathbf{y}}(u, v)$ interpolating the points $\overline{\mathbf{p}}_i$, i.e.,

$$\overline{\mathbf{y}}(u_i, v_i) = \overline{\mathbf{p}}_i, \tag{11}$$

yield the following conditions on the function $f(u, v)$:

$$f_i(u_i, v_i) = \frac{r_i}{|\mathbf{n}_i|}. \tag{12}$$

Moreover, we require that $\overline{\mathbf{y}}(u, v)$ touches the tangent planes $\overline{\alpha}_i$ at the given points $\overline{\mathbf{p}}_i$, i.e.,

$$\begin{array}{rcl} \overline{\mathbf{y}}_u(u_i, v_i) & = & \beta_{i1}\overline{\mathbf{t}}_{i1} + \beta_{i2}\overline{\mathbf{t}}_{i2}, \\ \overline{\mathbf{y}}_v(u_i, v_i) & = & \gamma_{i1}\overline{\mathbf{t}}_{i1} + \gamma_{i2}\overline{\mathbf{t}}_{i2}, \end{array} \tag{13}$$

which (for each of these identities) forms a system of four linear equations for three variables $f_u, \beta_{i1}, \beta_{i2}$ and $f_v, \gamma_{i1}, \gamma_{i2}$, respectively. However the equations in each system are dependent. This follows from the next dependency conditions:

$$\langle \widetilde{\mathbf{n}}, \overline{\mathbf{y}}_u \rangle = 0, \quad \langle \widetilde{\mathbf{n}}, \overline{\mathbf{y}}_v \rangle = 0, \tag{14}$$

where $\widetilde{\mathbf{n}} = (\mathbf{n}, |\mathbf{n}|) = (\mathbf{x}_u \times \mathbf{x}_v, |\mathbf{x}_u \times \mathbf{x}_v|)$ is an isotropic vector, see [8] for more details. So we can choose three arbitrary equations in each system, e.g., the first three of them, and solve

$$\begin{array}{rcl} \mathbf{y}_u(u_i, v_i) \cdot \mathbf{m}_i & = & 0, \\ \mathbf{y}_v(u_i, v_i) \cdot \mathbf{m}_i & = & 0, \end{array} \tag{15}$$

where $\mathbf{m}_i = \mathbf{t}_{i1} \times \mathbf{t}_{i2}$. Hence we arrive at

$$\begin{aligned} f_u^i(u_i, v_i) &= -\frac{(\mathbf{x}_u(u_i, v_i) + f_i(u_i, v_i)\mathbf{n}_u(u_i, v_i)) \cdot \mathbf{m}_i}{\mathbf{n}_i \cdot \mathbf{m}_i}, \\ f_v^i(u_i, v_i) &= -\frac{(\mathbf{x}_v(u_i, v_i) + f_i(u_i, v_i)\mathbf{n}_v(u_i, v_i)) \cdot \mathbf{m}_i}{\mathbf{n}_i \cdot \mathbf{m}_i}. \end{aligned} \tag{16}$$

After constructing a rational function $f(u, v)$ satisfying conditions (12) and (16), the corresponding envelope surface associated to $\overline{\mathbf{y}}(u, v)$ smoothly joins the given n spheres at the prescribed points and with the considered tangent planes.

The rational parametrization of the second branch $\mathbf{x}^-(u, v)$ of the envelope surface can be then easily computed as

$$\mathbf{x}^-(u, v) = \mathbf{x}^+ + 2f\frac{(\mathbf{y}_u \times \mathbf{y}_v) \cdot (\mathbf{x}_u \times \mathbf{x}_v)}{|\mathbf{x}_u \times \mathbf{x}_v|^2}(\mathbf{y}_u \times \mathbf{y}_v), \tag{17}$$

see Fig. 2 (bottom, left).

To sum up, the whole skinning element is composed of two branches of the envelope surface $\mathbf{x}^{\pm}(u, v)$, remaining parts of spheres and parts of canal surfaces whose rational descriptions can be obtained by rotating the boundary curves $\mathbf{x}_i(u)$ of $\mathbf{x}(u, v)$ around the tangents of the boundary curves $\mathbf{y}_i(u)$ of $\mathbf{y}(u, v)$, see Fig. 2 (bottom, right).

In particular, the construction of the skinning surface which smoothly joins 3 spheres is trivial – we construct a planar MST. For $n = 4$ we start with choosing tangent vectors $\overline{\mathbf{t}}_i$ at the points $\overline{\mathbf{p}}_i$, see (8). Next we find the points $\mathbf{q}_i$ and the associated tangent vectors $\mathbf{u}_{i1}$ and $\mathbf{u}_{i2}$ at these points and construct the Ferguson patch $\mathbf{x}(u, v)$, $u, v \in [0, 1]$, interpolating C^1 Hermite data $\mathbf{q}_i, \mathbf{u}_{i1}, \mathbf{u}_{i2}$. The next step is to construct the lifting function $f(u, v)$ as a one-dimensional Ferguson patch interpolating values (12) and (16), cf. Example 4.1.

4 Computed examples

In this section we present the designed method on two particular examples.

Example 4.1 Consider 4 spheres given by

$$\overline{\mathbf{p}}_1 = (0, 0, 0, 1), \quad \overline{\mathbf{p}}_2 = (8, 1, 1, 1), \quad \overline{\mathbf{p}}_3 = (6, 7, 0, 2), \quad \overline{\mathbf{p}}_4 = (0, 8, 1, 2), \tag{18}$$

see Fig. 3 (left). We start with setting tangent planes at $\overline{\mathbf{p}}_i$, i.e.,

$$\begin{array}{lll} \overline{\alpha}_1: & \overline{\mathbf{t}}_{11} = \overline{\mathbf{p}}_2 - \overline{\mathbf{p}}_1, & \overline{\mathbf{t}}_{12} = \overline{\mathbf{p}}_4 - \overline{\mathbf{p}}_1, \\ \overline{\alpha}_2: & \overline{\mathbf{t}}_{21} = \overline{\mathbf{p}}_2 - \overline{\mathbf{p}}_1, & \overline{\mathbf{t}}_{22} = \overline{\mathbf{p}}_3 - \overline{\mathbf{p}}_2, \\ \overline{\alpha}_3: & \overline{\mathbf{t}}_{31} = \overline{\mathbf{p}}_3 - \overline{\mathbf{p}}_4, & \overline{\mathbf{t}}_{32} = \overline{\mathbf{p}}_3 - \overline{\mathbf{p}}_2, \\ \overline{\alpha}_4: & \overline{\mathbf{t}}_{41} = \overline{\mathbf{p}}_3 - \overline{\mathbf{p}}_4, & \overline{\mathbf{t}}_{42} = \overline{\mathbf{p}}_4 - \overline{\mathbf{p}}_1 \end{array} \tag{19}$$

 ISBN: 978-84-608-6082-2

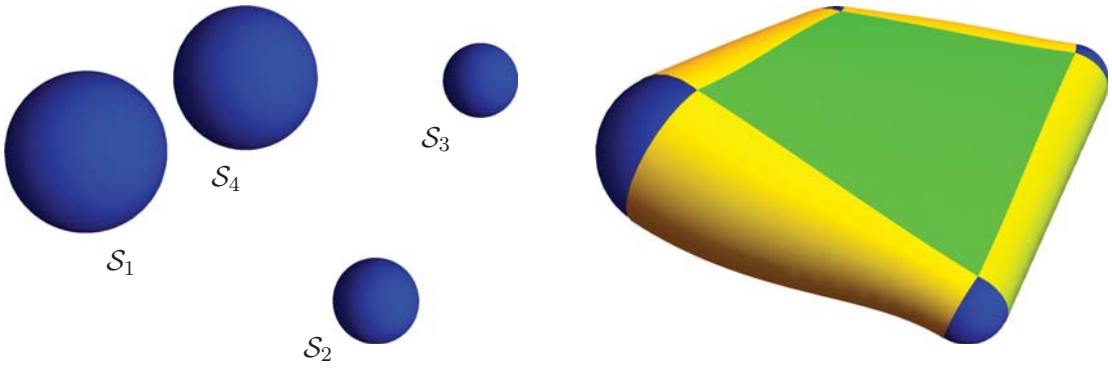


Figure 3: An illustration of the construction of the rational skinning surface between four given spheres from Example 4.1.

and compute the corresponding points $\mathbf{q}_i$ and normal vectors $\mathbf{n}_i$ on the associated envelope, cf. (9) and (10). Next we can choose the tangent vectors $\mathbf{u}_{i1}$, $\mathbf{u}_{i2}$, e.g. by projecting $\mathbf{t}_{i1}$, $\mathbf{t}_{i2}$ to the plane given by the point $\mathbf{q}_i$ and normal vector $\mathbf{n}_i$, i.e,

$$\mathbf{u}_{i1} = \mathbf{t}_{i1} - \frac{\mathbf{t}_{i1} \cdot \mathbf{n}_i}{|\mathbf{n}_i|^2}\mathbf{n}_i, \quad \mathbf{u}_{i2} = \mathbf{t}_{i2} - \frac{\mathbf{t}_{i2} \cdot \mathbf{n}_i}{|\mathbf{n}_i|^2}\mathbf{n}_i. \tag{20}$$

Then we construct a Ferguson surface $\mathbf{x}(u, v)$ interpolating points $\mathbf{q}_i$ and tangent vectors $\mathbf{u}_{i1}$ and $\mathbf{u}_{i2}$ and compute the lifting function $f(u, v)$ as a one dimensional Ferguson surface interpolating (12) and (16). Finally we compute the corresponding MST $\overline{\mathbf{y}}$ in form (7) and the second branch of the envelope (17). Of course, the boundary of the skin are rational canal surfaces obtained by rotating $\mathbf{x}(u, 0)$, $\mathbf{x}(u, 1)$, $\mathbf{x}(0, u)$, $\mathbf{x}(1, u)$ around $\mathbf{y}'(u, 0)$, $\mathbf{y}'(u, 1)$, $\mathbf{y}'(0, u)$, $\mathbf{y}'(1, u)$, respectively, see Fig. 3 (right).

Example 4.2 Consider a system of 10 spheres $\mathcal{S}_i$, cf. Fig 4 (left). We construct a branched skinning surface $\mathcal{S}$ of the system Σ such that the spheres $\mathcal{S}_1, \mathcal{S}_5, \mathcal{S}_6, \mathcal{S}_{10}$ will be joined by one suitable skinning element, cf. Section 3. Hence we can start with constructing this element but in this special case we choose the tangent vectors at $\overline{\mathbf{p}}_i$ more intentionally to enable simple joining with the remaining parts of the constructed skin. Consequently, we prescribe on the spheres $\mathcal{S}_1$, $\mathcal{S}_5$, $\mathcal{S}_6$ and $\mathcal{S}_{10}$ suitable circles and employ any arbitrary linear skinning method to the two system of spheres $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\}$ and $\{\mathcal{S}_6, \mathcal{S}_7, \mathcal{S}_8, \mathcal{S}_9, \mathcal{S}_{10}\}$, see Fig 4 (right).

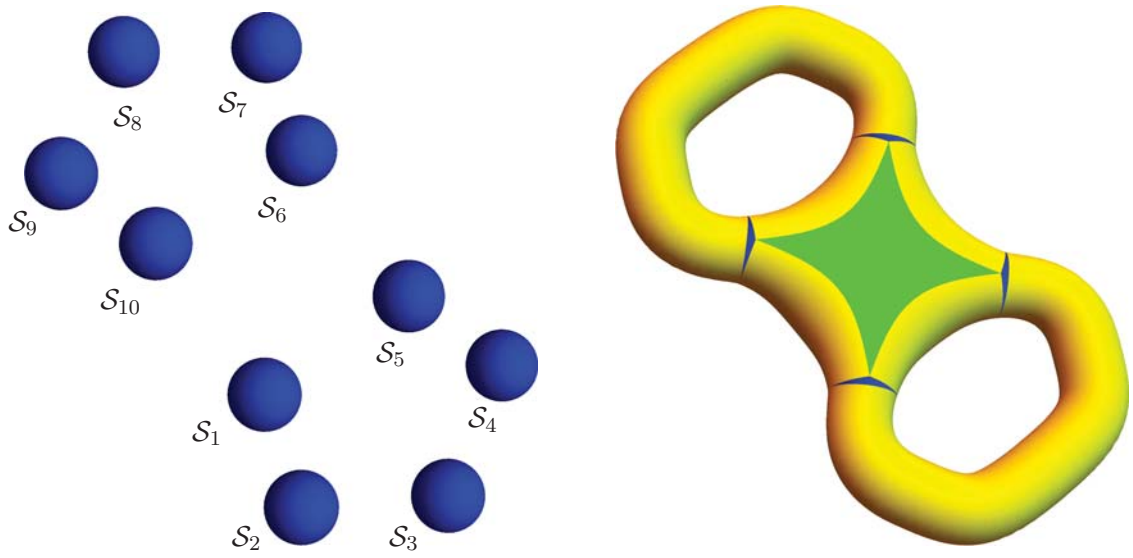


Figure 4: Construction of the rational branched skin of ten given spheres from Example 4.2.

5 Conclusion

This paper was devoted to the construction of branched skins of systems of spheres. Especially we focused on the parts which correspond to the envelopes of two-parametric systems of spheres. We presented how the so-called rational envelope surfaces can be used for solving this problem when skins with rational descriptions are required. The functionality of the presented method was presented on some examples.

Acknowledgements

The authors are supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

[1] B. Bastl, J. Kosinka, and M. Lávička, *Simple and branched skins of systems of circles and convex shapes*, Graphical Models, 78 (2015), pp. 1 – 9.

[2] M. Bizzarri and M. Lávička, *Parameterizing rational offset canal surfaces via rational contour curves*, Computer-Aided Design, 45 (2013), pp. 342–350.

[3] M. Bizzarri, M. Lávička, and J. Kosinka, *Medial axis transforms yielding rational envelopes*, Submitted to Computer Aided Geometric Design (under revision), (2015).

[4] H. Choi, C. Han, H. Moon, K. Roh, and N. S. Wee, *Medial axis transform and offset curves by Minkowski Pythagorean hodograph curves*, Computer-Aided Design, 31 (1999), pp. 59–72.

[5] M. Dohm and S. Zube, *The implicit equation of a canal surface*, J. Symb. Comput., 44 (2009), pp. 111–130.

[6] G. Farin, *Curves and surfaces for CAGD: A practical guide*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[7] J. Kosinka and B. Jüttler, *MOS surfaces: Medial surface transforms with rational domain boundaries*, in The Mathematics of Surfaces XII, vol. 4647 of Lecture Notes in Computer Science, Springer, 2007, pp. 245–262.

[8] J. Kosinka and M. Lávička, *A unified Pythagorean hodograph approach to medial axis transform and offset approximation*, Journal of Computational and Applied Mathematics, 235 (2011), pp. 3413–3424.

[9] R. Kunkli and M. Hoffmann, *Skinning of circles and spheres*, Computer Aided Geometric Design, 27 (2010), pp. 611–621.

[10] G. Landsmann, J. Schicho, and F. Winkler, *The parametrization of canal surfaces and the decomposition of polynomials into a sum of two squares*, Journal of Symbolic Computation, 32 (2001), pp. 119–132.

[11] H. Moon, *Minkowski Pythagorean hodographs*, Computer Aided Geometric Design, 16 (1999), pp. 739–753.

[12] M. Peternell and H. Pottmann, *A Laguerre geometric approach to rational offsets*, Computer Aided Geometric Design, 15 (1998), pp. 223–249.

[13] H. Pottmann, *Rational curves and surfaces with rational offsets*, Computer Aided Geometric Design, 12 (1995), pp. 175–192.

[14] H. Pottmann and M. Peternell, *Applications of Laguerre geometry in CAGD*, Computer Aided Geometric Design, 15 (1998), pp. 165–186.

[15] G. Slabaugh, B. Whited, J. Rossignac, T. Fang, and G. Unal, *3D ball skinning using PDEs for generation of smooth tubular surfaces*, Computer-Aided Design, 42 (2010), pp. 18–26.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An epidemic model for cholera with treatment through quarantine

Ana P. Lemos-Paião[1], Cristiana J. Silva[1] and Delfim F. M. Torres[1]

[1] *Center for Research and Development in Mathematics and Applications (CIDMA)*
Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

emails: anapaiao@ua.pt, cjoaosilva@ua.pt, delfim@ua.pt

Abstract

We propose a model for cholera with treatment through quarantine. We compute the disease-free equilibrium (DFE) and the basic reproduction number R_0. We also determine in which conditions the DFE is locally asymptotically stable. A numerical simulation of the cholera outbreak in the Department of Artibonite (Haiti), in 2010, is suggested and we prove that the number of infected individuals decreases when quarantine is applied. Finally, an optimal control problem, whose goal is to obtain a successful treatment through quarantine, is proposed.

*K*eywords: Cholera, SIQRB model, basic reproduction number, disease-free equilibrium, local stability, numerical simulation, optimal control.

*M*SC 2010: 34C60, 49K15, 92D30.

1 Introduction

Cholera is a bacterial disease provoked by the bacterium *Vibrio cholerae*, which lives in an aquatic organism. Therefore, the ingestion of contaminated water can cause cholera outbreaks, as John Snow proved, in 1854 [7]. This is a way of transmission of the disease, but there are others. For example, susceptible individuals can become infected if they contact with infected individuals. If individuals are at an increased risk of infection, they can transmit the disease to other persons that live with them by reflecting food preparation or using water storage containers [7]. An individual can be infected without or with symptoms. Some symptoms are watery diarrhoea, vomiting and leg cramps. If an infected individual

does not have treatment becomes dehydrated, suffering of acidosis and circulatory collapse. This situation can lead to death, within 12-24h [4, 7]. Some studies and experiments suggest that a recovered individual can be immune to the disease during a period of 3 to 10 years. Recently researches suggest that the immunity can be lost after a period of weeks to months [5, 7]. Between 2007 and 2011, there were cholera outbreaks in some countries of Africa and Asia, namely in Angola, Haiti and Zimbabwe [7]. In Haiti, the first cases of cholera happened in Artibonite Department on 14^{th} October 2010 and the disease propagated along the Artibonite river and reached several departments of Haiti. Only within one month, all departments had reported cases in rural areas and places without good conditions of public health [9]. Under some assumptions, we show that our model describes well such reality.

2 Model Formulation

We study a model that is based on [4, 5]. Our model divides the population density $N(t)$ into four different classes: $S(t)$, $I(t)$, $Q(t)$ and $R(t)$. The classes $S(t)$, $I(t)$, $Q(t)$ and $R(t)$ represent the number of individuals that are susceptible, infected with symptoms, in quarantine and recovered at time t, respectively. Furthermore, we will consider a class $B(t)$ that reflects the bacterial concentration at time t. We assume that there is a positive recruitment rate Λ into the susceptible class $S(t)$ and there is a positive natural death rate μ, for all time t. Susceptible individuals can become infected with cholera at rate $\frac{\beta B(t)}{\kappa+B(t)}$ that is dependent of time t. Note that $\beta > 0$ is the ingestion rate of the bacteria through contaminated sources, κ is the half saturation constant of the bacteria population and $\frac{B(t)}{\kappa+B(t)}$ is the possibility of an infected individual to have the disease with symptoms given a contact with contaminated sources. Any recovered individual can lose immunity at rate ω and therefore become susceptible again. The infected individuals can accept to be in quarantine during a period of time. During this time they are isolated and subject to a proper medication, at rate δ. The quarantined individuals can recover at rate ε. The disease-related death rates associated to the individuals that are infected and in quarantine are α_1 and α_2, respectively. Each infected individual contributes to the increase of the bacterial concentration at rate η. On the other side, the bacterial concentration can decrease at mortality rate d. These assumptions are translated in the following mathematical model:

$$\begin{cases} S'(t) = \Lambda - \dfrac{\beta B(t)}{\kappa + B(t)} S(t) + \omega R(t) - \mu S(t) \\ I'(t) = \dfrac{\beta B(t)}{\kappa + B(t)} S(t) - (\delta + \alpha_1 + \mu) I(t) \\ Q'(t) = \delta I(t) - (\varepsilon + \alpha_2 + \mu) Q(t) \\ R'(t) = \varepsilon Q(t) - (\omega + \mu) R(t) \\ B'(t) = \eta I(t) - dB(t). \end{cases} \tag{1}$$

3 Analysis of the Model

The density population is given by $N(t) = S(t) + I(t) + Q(t) + R(t)$ and, consequently, we have that

$$N'(t) = S'(t) + I'(t) + Q'(t) + R'(t) = \Lambda - \mu N(t) - \alpha_1 I(t) - \alpha_2 Q(t) \leq \Lambda - \mu N(t).$$

Using a standard comparison theorem (see [3]) one can easily show that $N(t) \leq \frac{\Lambda}{\mu}$ if $N(0) \leq \frac{\Lambda}{\mu}$. Thus, the region Ω given by

$$\Omega = \left\{ (S(t), I(t), Q(t), R(t), B(t)) \in \mathbb{R}^5_+ : N(t) \leq \frac{\Lambda}{\mu} \right\}$$

is positively invariant. Hence, it is sufficient to consider the dynamics of the flow generated by (1) in Ω. In this region, the model is epidemiologically and mathematically well posed in the sense of [2]. In other words, every solution of the model (1) with initial conditions in Ω remains in Ω for all $t > 0$.

Proposition 1 (Disease-free equilibrium) *The disease-free equilibrium of model* (1) *is given by* $E^0 = (S^0, I^0, Q^0, R^0, B^0) = \left(\frac{\Lambda}{\mu}, 0, 0, 0, 0\right)$.

To compute the basic reproduction number R_0 we allow the approach of [4, 8].

Proposition 2 (Basic reproduction number) *The basic reproduction number of model* (1) *is given by*

$$R_0 = \frac{\beta \Lambda \eta}{\mu \kappa d(\delta + \alpha_1 + \mu)}.$$

Furthermore, we have the following theorem.

Theorem 1 (Local asymptotic stability of DFE) *The disease-free equilibrium* E_0 *of model* (1) *is*

1. *Locally asymptotically stable, if* $\beta \Lambda \eta < \mu \kappa d(\delta + \alpha_1 + \mu)$;
2. *Unstable, if* $\beta \Lambda \eta > \mu \kappa d(\delta + \alpha_1 + \mu)$.

Moreover, if $\beta \Lambda \eta = \mu \kappa d(\delta + \alpha_1 + \mu)$, *then a critical case occurs.*

4 Numerical Simulations

We can observe that when $\omega = \delta = \varepsilon = 0$, we obtain a sub-model of (1) given by

$$\begin{cases} S'(t) = \Lambda - \dfrac{\beta B(t)}{\kappa + B(t)} S(t) - \mu S(t) \\ I'(t) = \dfrac{\beta B(t)}{\kappa + B(t)} S(t) - (\alpha_1 + \mu) I(t) \\ B'(t) = \eta I(t) - dB(t). \end{cases} \tag{2}$$

For a suitable choice of the parameters, the sub-model (2) can be an approximation of what happened in the Department of Artibonite (Haiti), since 1st November 2010 until 1st May 2011 [9]. In order to exist treatment for the infected individuals and, consequently, recovery, we have to suppose that $\delta, \varepsilon > 0$ and that $\omega \geq 0$. Considering the values of Table 1, we

Parameter	**Description**	**Value**	**Reference**
Λ	Recruitment rate	0.014 (day^{-1})	[1]
μ	Natural death rate	0.014 (day^{-1})	[1]
β	Ingestion rate	1.2 (day^{-1})	[1]
κ	Half saturation constant	10^6 (cell/ml)	[6]
ω	Immunity waning rate	0.4/365 (day^{-1})	[5]
δ	Quarantine rate	0.05 (day^{-1})	Assumed
ε	Recovery rate	0.2 (day^{-1})	[4]
α_1	Death rate (infected)	0.015 (day^{-1})	[4]
α_2	Death rate (quarantined)	0.0001 (day^{-1})	[4]
η	Shedding rate (infected)	10 (cell/ml day^{-1} person^{-1})	[1]
d	Bacteria death rate	0.33 (day^{-1})	[1]
$S(0)$	Susceptible individuals at $t = 0$	8000 (person)	Assumed
$I(0)$	Infected individuals at $t = 0$	1700 (person)	WHO
$Q(0)$	Quarantined individuals at $t = 0$	0 (person)	Assumed
$R(0)$	Recovered individuals at $t = 0$	0 (person)	Assumed
$B(0)$	Bacterial concentration at $t = 0$	10^4 (cell/ml)	Assumed

Table 1: Model parameter values.

have the graph of real data and its approximation. In Figure 1 we can observe the function $I(t)$, $t \in [0, 182]$, when there is a quarantine procedure.

5 Optimal Control Problem

In order to minimize the number of infected individuals and the bacterial concentration, as well as the cost of interventions associated to quarantine, we add to model (1) two control

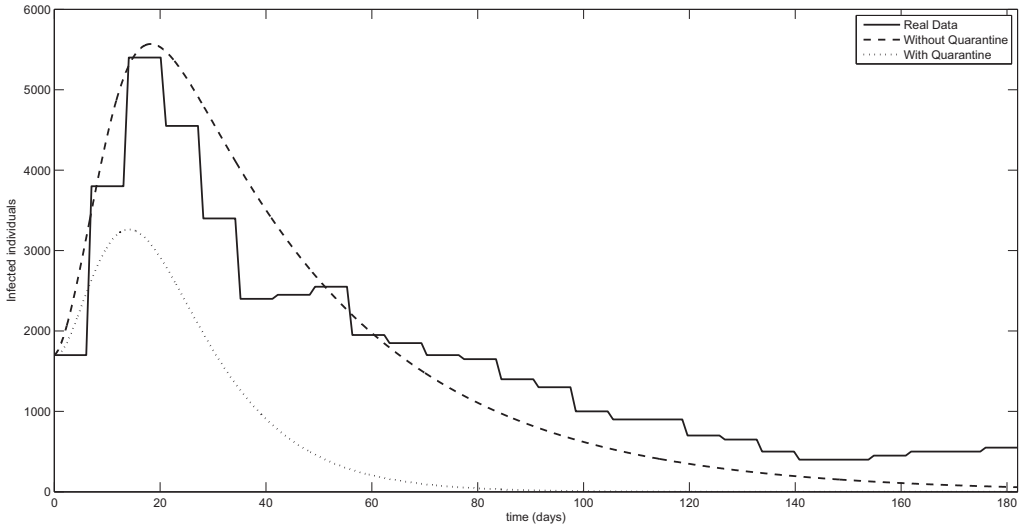


Figure 1: Numerical Simulation of real data of Artibonite with and without quarantine.

variables that promote the treatment through quarantine. The controls u_1 and u_2 reflect, respectively, the effort that is done for infected individuals to move to quarantine and for keeping it. Therefore, we obtain a model given by

$$\begin{cases} S'(t) = \Lambda - \dfrac{\beta B(t)}{\kappa + B(t)} S(t) + \omega R(t) - \mu S(t) \\ I'(t) = \dfrac{\beta B(t)}{\kappa + B(t)} S(t) - \delta u_1(t) I(t) - (\alpha_1 + \mu) I(t) \\ Q'(t) = \delta u_1(t) I(t) - \varepsilon u_2(t) Q(t) - (\alpha_2 + \mu) Q(t) \\ R'(t) = \varepsilon u_2(t) Q(t) - (\omega + \mu) R(t) \\ B'(t) = \eta I(t) - dB(t). \end{cases}$$

As we intend to minimize the number of infected individuals, the bacterial concentration and the cost of the measures associated to each control, from 1st November 2010 to 1st May 2011 [9], our goal is minimize the functional given by

$$J(u_1(.), u_2(.)) = \int_0^T \left(I(t) + B(t) + C_1 u_1^2(t) + C_2 u_2^2(t) \right),$$

where $T = 182$ days and C_i is the measure of cost associated to the control u_i, $i = 1, 2$.

Acknowledgements

This research was supported by the Portuguese Foundation for Science and Technology (FCT) within projects UID/MAT/04106/2013 (CIDMA) and PTDC/EEI-AUT/2933/2014 (TOCCATA). Lemos-Paião is also supported by the Ph.D. fellowship PD/BD/114184/2016; Silva by the post-doc fellowship SFRH/BPD/72061/2010.

References

[1] F. Capone, V. De Cataldis and R. De Luca, *Influence of diffusion on the stability of equilibria in a reaction–diffusion system modeling cholera dynamic*, J. Math. Biol. **71** (2015) 1107–1131.

[2] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev. **42** (2000) 599–653.

[3] V. Lakshmikantham, S. Leela and A. A. Martynyuk, *Stability Analysis of Nonlinear Systems*, Marcel Dekker, Inc., New York and Basel, 1989.

[4] A. Mwasa and J. M. Tchuenche, *Mathematical analysis of a cholera model with public health interventions*, Bull. Math. Biol. **105** (2011) 190–200.

[5] R. L. M. Neilan, E. Schaefer, H. Gaff, K. R. Fister and S. Lenhart, *Modeling Optimal Intervention Strategies for Cholera*, Bull. Math. Biol. **72** (2010) 2004–2018.

[6] R. P. Sanches, C. P. Ferreira and R. A. Kraenkel, *The Role of Immunity and Seasonality in Cholera Epidemics*, Bull. Math. Biol. **73** (2011) 2916–2931.

[7] Z. Shuai, J. H. Tien and P. van den Driessche, *Cholera Models with Hyperinfectivity and Temporary Immunity*, Bull. Math. Biol. **74** (2012) 2423–2445.

[8] P. van den Driessche and J. Watmough, *Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission*, Math. Biosc. **180** (2002) 29–48.

[9] World Health Organization, Global Task Force on Cholera Control, Cholera Country Profile: Haiti, 18 May 2011, `http://www.who.int/cholera/countries/HaitiCountryProfileMay2011.pdf`.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An overview of canonical Euler splitting methods for nonlinear composite stiff evolution equations

Shoufu Li[1]

[1] *School of Mathematics and Computational Science, Xiangtan University, Xiangtan, 411105, China*

emails: lisf@xtu.edu.cn

Abstract

In previous papers, the author has constructed and studied canonical Euler splitting method (CES) and generalized canonical Euler splitting methods (GCES). Theoretical analysis and numerical experiments show that CES and GCES methods are universally applicable to general nonlinear composite stiff problems in evolution equations of various type, and can significantly improve the computing speed on the basis of computing quality assurance, whereas all the traditional operator splitting methods neither have such universal applicability, nor have the fast computing speed which can be compared with that of CES and GCES methods.

Key words: Canonical Euler splitting methods, nonlinear composite stiff problems, evolution equations.

1 Introduction

Splitting algorithm is an essential key technology for the numerical simulation of complex multi-physics processes and large-scale scientific computing, it has been widespread concerned by experts and scientific computing staffs at home and abroad.

The traditional operator splitting methods, such as sequential splitting, symmetrical weighted sequential splitting, Strang-Marchuk splitting and iterative splitting methods, have been widely used for solving evolution equations. However, these methods are designed based on the theory of linear operator and operator semigroups on Banach space, which can not be directly applied to general nonlinear differential equations and stiff problems. Furthermore, even for a traditional operator splitting method applied to a linear stiff system

ISBN: 978-84-608-6082-2

in ODEs, in case the related linear operators do not commute, the coefficient of the local splitting error may become extremely large and cause the splitting method losing theoretical basis and any practical value.

In recent years, through the use of local linearization hypothesis, Lie-derivative and other special approaches and techniques to improve the traditional operator splitting methods, scientists have constructed a series of new operator splitting methods, which can be effectively used to solve various specific nonlinear evolution equations encountered in the research of modern science and technology. As examples, splitting methods for solving nonlinear hyperbolic conservation laws we refer to [32, 36], splitting methods for solving unsteady convection-diffusion-reaction equations with nonlinear convective terms we refer to [31, 11, 37], with nonlinear reaction terms refer to [22, 12, 19, 16, 34, 29], with nonlinear diffusion terms refer to [30, 38, 35, 9, 2]. Furthermore, splitting methods for solving some special nonlinear delay differential equations (DDEs) and integro-differential equations (IDEs) can be found in [6, 10, 13, 1, 20, 4, 18].

In order to further overcome the drawbacks of traditional operator splitting methods mentioned above, in papers [23, 28], the author has constructed and studied canonical Euler splitting method (CES) and generalized canonical Euler splitting methods (GCES). Theoretical analysis and numerical experiments show that CES and GCES methods are universally applicable to general nonlinear composite stiff problems in evolution equations of various type, and can significantly improve the computing speed on the basis of computing quality assurance, whereas all the traditional operator splitting methods neither have such universal applicability, nor have the fast computing speed which can be compared with that of CES and GCES methods.

In this paper we firstly consider the initial value problem in Volterra functional differential equations (VFDEs) of the form

$$\begin{cases} y'(t) = f_1(t, y(t), y(\cdot)) + f_2(t, y(t), y(\cdot)), & t \in (0, T], \\ y(t) = \varphi(t), & t \in [-\tau, 0], \end{cases} \tag{1.1}$$

where $T > 0$, $\tau \in [0, +\infty]$ are constants, $\varphi \in \mathbf{C}_m[-\tau, 0]$ is a given initial function, $f_1, f_2 : [0, T] \times \mathbf{R}^m \times \mathbf{C}_m[-\tau, T] \to \mathbf{R}^m$ are given mappings, which satisfy the conditions

$$\begin{cases} \| f_1(t, u, \psi(\cdot)) - f_1(t, v, \psi(\cdot)) \| \leq \alpha_1 \| u - v \| & \forall t \in [0, T], u, v \in \mathbf{R}^m, \psi \in \mathbf{C}_m[-\tau, T], \quad (1.2a) \\ \| f_1(t, u, \psi(\cdot)) - f_1(t, u, \chi(\cdot)) \| \leq \beta_1 \max\limits_{-\tau \leq \xi \leq t} \| \psi(\xi) - \chi(\xi) \| & \forall t \in [0, T], u \in \mathbf{R}^m, \psi, \chi \in \mathbf{C}_m[-\tau, T], \quad (1.2b) \\ \langle f_2(t, u, \psi(\cdot)) - f_2(t, v, \psi(\cdot)), u - v \rangle \leq \overline{\alpha}_2 \| u - v \|^2 & \forall t \in [0, T], u, v \in \mathbf{R}^m, \psi \in \mathbf{C}_m[-\tau, T], \quad (1.2c) \\ \| f_2(t, u, \psi(\cdot)) - f_2(t, u, \chi(\cdot)) \| \leq \beta_2 \max\limits_{-\tau \leq \xi \leq t} \| \psi(\xi) - \chi(\xi) \| & \forall t \in [0, T], u \in \mathbf{R}^m, \psi, \chi \in \mathbf{C}_m[-\tau, T]. \quad (1.2d) \end{cases}$$

Here $\mathbf{R}^m$ denotes an m dimensional Euclidian space with the inner product $\langle \cdot, \cdot \rangle$ and the corresponding norm $\| \cdot \|$, for any given closed interval $\mathbf{I} \subset \mathbf{R}$, the symbol $\mathbf{C}_m(\mathbf{I})$ denotes

 ISBN: 978-84-608-6082-2

a Banach space consisting of all continuous mappings $x : \mathbf{I} \to \mathbf{R}^m$, on which the norm is defined by $\| x \|_\infty = \max_{t\in\mathbf{I}} \| x(t) \|$, in (1.2) $\overline{\alpha}_2$ is a one-sided Lipschitz constant, α_1, β_1 and β_2 are classical Lipschitz constants, and the parameters α_1, β_1, $\overline{\overline{\alpha}}_2 := \max\{\overline{\alpha}_2, 0\}$, β_2 and T are always assumed to be of moderate size.

Furthermore, we assume that the problem (1.1) has a unique true solution $y(t)$ which is slowly varying and stable with respect to the initial function $\varphi(t)$. Here the term "slowly varying" means that $y(t)$ and all its derivatives used in the study are continuous and satisfy

$$\left\| \frac{d^i y(t)}{dt^i} \right\| \le M_i, \qquad 0 \le t \le T. \tag{1.3}$$

with each constant M_i to be of moderate size.

In this paper, we shall always use the symbol $\mathcal{S}(\alpha_1, \beta_1, \overline{\alpha}_2, \beta_2, T)$ (abbr. $\mathcal{S}$) to denote the VFDE problem class consisting of all the problems of the form (1.1) satisfying (1.2), (1.3) and all the above mentioned assumption conditions.

For the very important special case where the mapping f does not depend on the past values of the true solution $y(t)$, the problem (1.1) degenerates to the initial value problem in ODEs

$$\begin{cases} y'(t) = f(t, y(t)) := f_1(t, y(t)) + f_2(t, y(t)), & t \in (0, T] \qquad (1.4a) \\ y(0) = \varphi_0, & \varphi_0 \in \mathbf{R}^m, \qquad (1.4b) \end{cases}$$

which satisfies the conditions

$$\begin{cases} \| f_1(t, u) - f_1(t, v) \| \le \alpha_1 \| u - v \| & \forall t \in [0, T], u, v \in \mathbf{R}^m, \qquad (1.5a) \\ \langle f_2(t, u) - f_2(t, v), u - v \rangle \le \overline{\alpha}_2 \| u - v \|^2 & \forall t \in [0, T], u, v \in \mathbf{R}^m. \qquad (1.5b) \end{cases}$$

with the parameters α_1, $\overline{\overline{\alpha}}_2 := \max\{\overline{\alpha}_2, 0\}$ and T to be of moderate size, and here we also assume that the problem (1.4) has a unique true solution $y(t)$ which is slowly varying (i.e., satisfies the inequality (1.3)) and stable with respect to the initial value φ_0. Thus we get a problem class $\mathcal{S}_0(\alpha_1, \overline{\alpha}_2, T)$ (abbr. $\mathcal{S}_0$) consisting of all the problems of the form (1.4) satisfying (1.5), (1.3) and all the aforementioned assumptions, which can be regarded as a sub-class of the problem class $\mathcal{S}$.

Note that problem (1.4) can also be regarded as a semi-discrete PDE problem.

For any given problem (1.1) (resp. (1.4)), using the sub-mappings f_1 and f_2, we can form two sub-problems, and it is easily seen from the above assumption conditions that the first sub-problem with the sub-mapping f_1 as right hand side function is always non-stiff or

 ISBN: 978-84-608-6082-2

mild stiff. but the second sub-problem with the sub-mapping f_2 as right hand side function may be non-stiff, mild stiff, stiff or strongly stiff. For the case where the second sub-problem is also non-stiff or mild stiff, we can use any one explicit canonical Runge-Kutta method for VFDEs (resp. Runge-Kutta method for ODEs) to solve the original problem (1.1) (resp. (1.4)) efficiently (cf.[26, 27]), and it is no need to use splitting methods. Thus we always focus on the case where the second sub-problem is stiff or strongly stiff, and for convenience we call the original problem (1.1) (resp. (1.4)) to be composite stiff.

In practice, the problems called stiff are diverse, and thus various definitions of stiffness are seen in the literature. For the concepts of stiffness and for how to assess whether a ODE or VFDE problem is stiff or not, we refer to [7, 21, 33, 14, 26]. For any stiff problem, it is required that the true solution of the problem is slowly varying (cf. [7]). However, we emphasize that for a composite stiff problem, we only require the true solution of the original problem slowly varying, do not require the true solutions of its sub-problems slowly varying, in other words, for the above mentioned two sub-problems, their true solutions are allowed to vary no matter how fast (cf.Examples 5.1 and 5.2 of paper [23]). To avoid confusion, here we would like to give a practical but not rigorous definition: a sub-problem of the composite stiff problem (1.1) (resp. (1.4)) is called to be mild stiff, stiff or strongly stiff if and only if the right hand side function of it satisfies a classical Lipschitz condition with respect to the second argument, where the minimum Lipschitz constant is a little large, large, or very large, respectively, and the true solution of the original problem (1.1) (resp. (1.4)) is slowly varying.

Note that the conditions (1.2) and (1.5) seem to be too severe. In practice, these conditions can be in general weakened that we only need these conditions to be satisfied in some neighbourhood of the true solution $y(t)$ of the problem (cf.[15, 7, 27]). Furthermore, from a practical point of view, there is no need to check the conditions (1.2) and (1.5) strictly.

2 CES and GCES methods

In paper [23], the author have constructed and studied canonical Euler splitting method (CES) for solving nonlinear composite stiff problems of the form (1.1)$\in \mathcal{S}(\alpha_1, \beta_1, \overline{\alpha}_2, \beta_2, T)$ and (1.4)$\in \mathcal{S}_0(\alpha_1, \overline{\alpha}_2, T)$. Quantitative stability, consistency and convergence theories of

 ISBN: 978-84-608-6082-2

CES method are established, and it is proved that CES method is quantitatively consistent and convergent of order 1. Here and in the rest of this paper, the term "quantitative stability" means that in the related stability inequalities, all the stability coefficients and the reciprocal of the maximum permitted step size are of moderate size, and the term "quantitative consistency and convergence" means that in the related consistency and convergence inequalities, all the error coefficients and the reciprocal of the maximum permitted step size are of moderate size.

For convenience of the reader, we present the details in the following.

(1) CES method is very applicable to solve various nonlinear composite stiff VFDE problems of the form (1.1) belonging to the problem class $\mathcal{S}$ (cf.Example 5.2 of [23]). In such situation, to advance any one time integration step $(t_n, \psi, y_1, y_2, \cdots, y_n) \to (t_{n+1}, \psi, y_1, y_2, \cdots, y_{n+1})$, CES method is carried out according to the following steps:

Step 1. decompose the problem (1.1) into two sub-problems, i.e., the non-stiff sub-problem in ODEs

$$\begin{cases} \bar{y}'(t) = f_1(t, \bar{y}(t), y^h(\cdot)), \quad t \in [t_n, t_{n+1}], \\ \bar{y}(t_n) = y_n \end{cases} \quad 2 \tag{2.1a}$$

and the stiff sub-problem in ODEs

$$\begin{cases} \widehat{y}'(t) = f_2(t, \widehat{y}(t), y^h(\cdot)), \quad t \in [t_n, t_{n+1}], \\ \widehat{y}(t_n) = \bar{y}_{n+1}, \end{cases} \tag{2.1b}$$

where the function $y^h(\cdot) \in \mathrm{C}_m[-\tau, t_{n+1}]$ can be determined in advance by the formula

$$y^h(t) = \mathit{\Pi}^h(t; \psi, y_1, y_2, \cdots, y_n), \quad -\tau \le t \le t_{n+1}. \tag{2.2}$$

Throughout this paper, we always assume that the time mesh $\Delta_h := \{t_i : i = 0|N\} \in \{\Delta_h\}$ satisfies $0 = t_0 < t_1 < \cdots < t_N = T$ with variable stepsize $h_i = t_{i+1} - t_i$ and $h = \max\limits_{0 \le i \le N-1} h_i$, $\psi \in \mathbf{C}_m[-\tau, 0]$ is an approximation to the initial function φ, $y_i \in \mathbf{R}^m$ $(i = 0|N)$ are approximations to $y(t_i)$, $y^h(t)$ is an approximation to $y(t)$, and $\mathit{\Pi}^h : \mathbf{C}_m[-\tau, 0] \times \mathbf{R}^{mn} \to \mathbf{C}_m[-\tau, t_{n+1}]$ denotes a piecewise Lagrangian interpolation operator which satisfies the canonical condition

$$\max_{-\tau \le t \le t_{n+1}} \| \mathit{\Pi}^h(t; \psi, y_1, \cdots, y_n) - \mathit{\Pi}^h(t; \chi, z_1, \cdots, z_n) \| \le C_\pi \max\{ \max_{1 \le i \le n} \| y_i - z_i \|, \max_{-\tau \le t \le 0} \| \psi(t) - \chi(t) \| \}$$
$$\forall \psi, \chi \in \mathbf{C}_m[-\tau, 0],\ y_i, z_i \in \mathbf{R}^m,\ i = 1, 2, \cdots, n, \tag{2.3}$$

 ISBN: 978-84-608-6082-2

with the canonical constant C_π to be of moderate size.

For the special case of CES method, we always assume that the degree of Lagrangian integration polynomials does not greater than 1 and thus we have $C_\pi = 1$. For more details of the above symbols and the canonical condition (2.8), we refer to [25, 26, 27].

Step 2. Using generalized explicit Euler method (GEE)

$$\bar{y}_{n+1} = y_n + h_n f_1(t_{n+1}, y_n, y^h(\cdot)) \tag{2.4a}$$

to solve the non-stiff sub-problem (2.1a), we thus get $\bar{y}_{n+1}$.

Step 3. Using generalized implicit Euler method (GIE)

$$\widehat{y}_{n+1} = \bar{y}_{n+1} + h_n f_2(t_{n+1}, \widehat{y}_{n+1}, y^h(\cdot)) \tag{2.4b}$$

to solve the stiff sub-problem (2.1b) to get $\widehat{y}_{n+1}$. Let $y_{n+1} = \widehat{y}_{n+1}$, and we thus complete the time integration step.

Here we emphasize that in both (2.4a) and (2.4b) the function $y^h(t)$ is determined by formula (2.2) in which the original back value y_n (rather than $\bar{y}_{n+1}$) is used. Furthermore, we also emphasize that for solving equation (2.4b) by simplified Newton iterations to find $\widehat{y}_{n+1}$, the iteration starting value should be chosen as the original bach value y_n rather than $\bar{y}_{n+1}$ because we have only assumed that the true solution $y(t)$ of the original problem (1.1) is slowly varying, whereas the true solution $\widehat{y}(t)$ of the subproblem (2.1b) may be varying very fast.

(2) CES method is very applicable to solve various nonlinear composite stiff ODE or semi-discrete PDE problems of the form (1.4) belonging to the problem class $\mathcal{S}_0$ (cf.Example 5.1 of [23]). In such situation, to advance any one time integration step $(t_n, y_n) \to (t_{n+1}, y_{n+1})$, CES method is carried out according to the following steps:

Step 1. decompose the problem (1.4) into two sub-problems, i.e., the non-stiff sub-problem in ODEs

$$\begin{cases} \bar{y}'(t) = f_1(t, \bar{y}(t)), \quad t \in [t_n, t_{n+1}], \\ \bar{y}(t_n) = y_n \end{cases} \tag{2.5a}$$

 ISBN: 978-84-608-6082-2

and the stiff sub-problem in ODEs

$$\begin{cases} \widehat{y}'(t) = f_2(t, \widehat{y}(t)), \quad t \in [t_n, t_{n+1}], \\ \widehat{y}(t_n) = \bar{y}_{n+1}. \end{cases} \tag{2.5b}$$

Step 2. Using generalized explicit Euler method (GEE)

$$\bar{y}_{n+1} = y_n + h_n f_1(t_{n+1}, y_n) \tag{2.6a}$$

to solve the non-stiff sub-problem (2.5a), we thus get $\bar{y}_{n+1}$.

Step 3. Using generalized implicit Euler method (GIE)

$$\widehat{y}_{n+1} = \bar{y}_{n+1} + h_n f_2(t_{n+1}, \widehat{y}_{n+1}) \tag{2.6b}$$

to solve the stiff sub-problem (2.5b), we get $\widehat{y}_{n+1}$. Let $y_{n+1} = \widehat{y}_{n+1}$, and we thus complete the integration step.

Here we also emphasize that for solving equation (2.6b) by simplified Newton iterations to find $\widehat{y}_{n+1}$, the iteration starting value should be chosen as the original back value y_n rather than $\bar{y}_{n+1}$.

(3) For a composite stiff problem, if the first sub-problem is mild stiff, but the second sub-problem is strongly stiff, then CES method can also be used to solve it, but in this special case, the time step size must be chosen more smaller (cf. Examples 5.3, 5.4 and 5.5 of [23]).

However, CES method has a disadvantage that for solving sub-problems, it limits the methods too strict. In more detail, for solving non-stiff and stiff subproblems, it only allows to use GEE and GIE methods, respectively. To overcome this drawback and further develop CES method, in paper [28] the author has proved that for solving a nonlinear composite stiff problem belonging to the class $\mathcal{S}$ or $\mathcal{S}_0$, if the subproblems of this problem is also slowly varying, then we can advance any one time integration step from t_n to t_{n+1} according to the aforementioned three steps with GEE and GIE methods replaced by the methods $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, and we thus call such a new splitting method generalized canonical Euler splitting method with abbreviation GCES, or more precisely, GCES($\mathcal{M}_1$,$\mathcal{M}_2$), where $\mathcal{M}_1$

denotes an arbitrarily given Runge-Kutta method, which is classically stable, classically consistent and classically convergent of order $p_1 \geq 1$ for solving non-stiff problems, $\mathcal{M}_2$ denotes an arbitrarily given Runge-Kutta method, which is B-stable, B-consistent and optimally B-convergent of order $p_2 \geq 1$ for solving nonlinear stiff problems. Furthermore, we have also proved that GCES methods are quantitatively stable, quantitatively consistent and convergent of order 1 for solving nonlinear composite stiff problems belonging to the class $\mathcal{S}$ or $\mathcal{S}_0$ with the aforementioned additional property.

3 Numerical experiments

Example 4.1 Consider the strongly nonlinear three-dimensional parabolic problem

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}(u^6\frac{\partial u}{\partial x}) + \frac{\partial}{\partial y}(u^6\frac{\partial u}{\partial y}) + \frac{\partial}{\partial z}(u^6\frac{\partial u}{\partial z}) + \varphi, & (x,y,z) \in \ \Omega,\ 0 < t \leq \pi/2, \\ u(x,y,z,0) = 0, & (x,y,z) \in \ \Omega, \\ u(x,y,z,t) = 0, & (x,y,z) \in \ \partial\Omega,\ 0 \leq t \leq \pi/2, \end{cases} \tag{4.1}$$

where $\Omega = \{(x,y,z) : 0 < x < 1,\ 0 < y < 10,\ 0 < z < 10\}$, $\partial\Omega$ denotes the Lipschitz continuous boundary of Ω,

$$\varphi(x,y,z,t) = 2\sin(\pi x)\sin(\frac{\pi y}{10})\sin(\frac{\pi z}{10})\cos t + \frac{51\pi^2}{50}u^7 - 6\pi^2 u^5 \sin^2 t(4\cos^2(\pi x)\sin^2(\frac{\pi y}{10})\sin^2(\frac{\pi z}{10})$$
$$+\frac{1}{25}\sin^2(\pi x)\cos^2(\frac{\pi y}{10})\sin^2(\frac{\pi z}{10}) + \frac{1}{25}\sin^2(\pi x)\sin^2(\frac{\pi y}{10})\cos^2(\frac{\pi z}{10})).$$

This problem has a unique true solution

$$u(x,y,z,t) = 2\sin(\pi x)\sin(\frac{\pi y}{10})\sin(\frac{\pi z}{10})\sin t, \quad x,y,z \in \ \Omega, \quad 0 \leq t \leq \frac{\pi}{2}. \tag{4.2}$$

Using uniform space mesh

$$\{(x_i,y_j,z_k) \ | \ x_i = ih_x, y_j = jh_y, z_k = kh_z,\ i,j,k = 0,1,\cdots,100,\ h_x = 0.01, \quad h_y = h_z = 0.1\},$$

we get the following semi-discrete problem by difference method

$$\begin{cases} \frac{du_{ijk}}{dt} = \frac{(u^6_{i+1,j,k} + u^6_{ijk})(u_{i+1,j,k} - u_{ijk}) - (u^6_{ijk} + u^6_{i-1,j,k})(u_{ijk} - u_{i-1,j,k})}{2h_x^2} \\ \qquad + \frac{(u^6_{i,j+1,k} + u^6_{ijk})(u_{i,j+1,k} - u_{ijk}) - (u^6_{ijk} + u^6_{i,j-1,k})(u_{ijk} - u_{i,j-1,k})}{2h_y^2} \\ \qquad + \frac{(u^6_{i,j,k+1} + u^6_{ijk})(u_{i,j,k+1} - u_{ijk}) - (u^{q_1}_{ijk} + u^{q_1}_{i,j,k-1})(u_{ijk} - u_{ijk-1})}{2h_z^2} + \varphi_{i,j,k}(t), \\ u_{ijk}(0) = 0, \end{cases} \tag{4.3}$$

 ISBN: 978-84-608-6082-2

where $0 < t \leq \frac{\pi}{2}$, $i, j, k = 1, 2, \cdots, 99$, $\varphi_{i,j,k}(t) = \varphi(x_i, y_j, z_k, t)$, the unknown functions $u_{i,j,k} = u_{i,j,k}(t)$ are approximations to $u(x_i, y_j, z_k, t)$, and note that $u_{0,j,k}(t) = u_{100,j,k}(t) = u_{i,0,k}(t) = u_{i,100,k}(t) = u_{i,j,0}(t) = u_{i,j,100}(t) = 0$.

For each time integration step from t_n to t_{n+1}, we decompose the semi-discrete problem (4.3) into two sub-problems, i.e., the mild-stiff sub-problem

$$\begin{aligned}\frac{d\bar{u}_{ijk}}{dt} &= \frac{(\bar{u}^6_{i,j+1,k}+\bar{u}^6_{ijk})(\bar{u}_{i,j+1,k}-\bar{u}_{ijk})-(\bar{u}^6_{ijk}+\bar{u}^6_{i,j-1,k})(\bar{u}_{ijk}-\bar{u}_{i,j-1,k})}{2h_y^2} \\ &+ \frac{(\bar{u}^6_{i,j,k+1}+\bar{u}^6_{ijk})(\bar{u}_{i,j,k+1}-\bar{u}_{ijk})-(\bar{u}^6_{ijk}+\bar{u}^6_{i,j,k-1})(\bar{u}_{ijk}-\bar{u}_{i,j,k-1})}{2h_z^2} + \varphi_{i,j,k}(t)\end{aligned} \tag{4.4a}$$

and the strongly stiff sub-problem

$$\frac{d\tilde{u}_{ijk}}{dt} = \frac{(\tilde{u}^6_{i+1,jk} + \tilde{u}^6_{ijk})(\tilde{u}_{i+1,jk} - \tilde{u}_{ijk}) - (\tilde{u}^6_{ijk} + \tilde{u}^6_{i-1,jk})(\tilde{u}_{ijk} - \tilde{u}_{i-1,jk})}{2h_x^2}. \tag{4.4b}$$

Then we solve this semi-discrete problem by our CES method with variable time stepsize τ_n determined by the empirical formula based on linear stability theory

$$\tau_n = \frac{(\min\{h_y, h_z\})^2}{c \max\{1, u^6_{max}(t_n)\}}, \qquad u_{max}(t) = \max_{i,j,k} u_{i,j,k}(t), \tag{4.5}$$

where $c \geq 2$ is an adjustable parameter, and we let $c = 4$. We find that the L2-norm of the global error of the numerical solution with respect to the true solution $u(x, y, z, t)$ of the original PDE problem (4.1) is $E_{L_2} = 1.368975 \times 10^{-3}$, and the computation only takes time 11 hours 18 minutes 7.296 seconds.

Note that there are nearly 10^6 space grids and 10^6 unknown functions $u_{ijk}(t), 0 \leq t \leq \frac{\pi}{2}$, computed with time stepsize τ_n gradually decreased from 2.5×10^{-3} to 3.911212×10^{-5}.

Note also that in this example we always perform our codes in serial on the same Dell OptiPlex 755 computer and always require Newton iteration error $\varepsilon < 10^{-10}$ to stop the simplified Newton iterations. If we perform our codes in parallel, then the computation speed can be certainly further improved significantly.

Therefore, we can conclude that CES method is very applicable for solving strongly nonlinear composite stiff semi-discrete multi-dimensional parabolic problems, it can significantly improve the computing speed on the basis of ensuring the computing quality.

Example 4.2 For the numerical simulation of radiation driven spherically symmetric single-

 ISBN: 978-84-608-6082-2

temperature implosion compression process, in Eulerian coordinates, its mathematical model is

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial(\rho u)}{\partial x} + \dfrac{2\rho u}{x} = 0, & (4.6a) \\ \dfrac{\partial(\rho u)}{\partial t} + \dfrac{\partial(\rho u^2 + p)}{\partial x} + \dfrac{2\rho u^2}{x} = 0, & (4.6b) \\ \dfrac{\partial(\rho E)}{\partial t} + \dfrac{\partial((\rho E + p)u)}{\partial x} + \dfrac{2(\rho E + p)u}{x} = \dfrac{\partial}{\partial x}(K\dfrac{\partial T}{\partial x}) + \dfrac{2K\frac{\partial T}{\partial x}}{x}, & (4.6c) \\ \dfrac{\partial \rho\nu}{\partial t} + \dfrac{\partial(\rho\nu u)}{\partial x} + \dfrac{2\rho\nu u}{x} = 0, & (4.6d) \\ p = (\gamma - 1)\rho e, \qquad e = c_v T, \qquad E = e + \dfrac{1}{2}u^2, & (4.6e) \end{cases}$$

where the symbols $\rho,\ u,\ e,\ E,\ p,\ T$ and ν denote the fluid density, velocity, specific internal energy, total energy, pressure, temperature and volume fraction of deuterium-tritium (DT) , respectively, all of them are functions of (x,t) with $0 \le x \le 5R$ and $t \ge 0$, $R = a + b + c$ denotes the radius of the target sphere (cf. Figure 4.1), for definiteness and simplicity, let $a = 114,\ b = 0,\ c = 14$, and we thus have $R = 128$, K, γ and c_v denote thermal conductivity, adiabatic index of gas and isochoric specific heat, respectively, which are dependent on materials. For DT, we let $\gamma = 5/3,\ c_v = 98.8$ and for CH, let $\gamma = 5/3,\ c_v = 86$, since K is a complex function of T and ρ, here we do not describe it in detail.

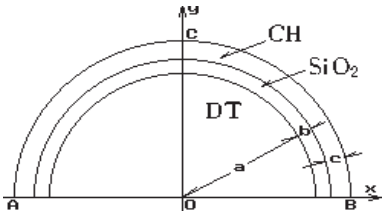


Figure 4.1 target sphere structure

At the left and right ends of the interval $[0, 5R]$, we use symmetric and compactly supported boundary conditions, respectively. At the initial time $t = 0$, we let $u(x,0) = 0,\ T(x,0) = 3 \times 10^{-4}$ for $x \in [0,\ 5R]$, and let $\rho(x,0) = 0.0018,\ 1.05,\ 10^{-5}$ for $x \in [0,a],\ [a,R],\ [R,5R]$ respectively. Furthermore, we assume that there is a radiation source on the interval $[R,\ 5R]$, and the temperature of the radiation source is determined by the formula

$$T_r = \begin{cases} 3 \times 10^{-4} + (2 - 3 \times 10^{-4})t, & 0 \le t \le 1, \\ 2, & t > 1. \end{cases}$$

 ISBN: 978-84-608-6082-2

Problem (4.6) can be decomposed into two subproblems, i.e., the fluid flow subproblem

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial(\rho u)}{\partial x} + \dfrac{2\rho u}{x} = 0, & (4.7a) \\ \dfrac{\partial(\rho u)}{\partial t} + \dfrac{\partial(\rho u^2 + p)}{\partial x} + \dfrac{2\rho u^2}{x} = 0, & (4.7b) \\ \dfrac{\partial(\rho E)}{\partial t} + \dfrac{\partial((\rho E + p)u)}{\partial x} + \dfrac{2(\rho E + p)u}{x} = 0, & (4.7c) \\ \dfrac{\partial \rho v}{\partial t} + \dfrac{\partial(\rho v u)}{\partial x} + \dfrac{2\rho v u}{x} = 0, & (4.7d) \\ p = (\gamma - 1)\rho e, \qquad E = e + \dfrac{1}{2}u^2, & (4.7e) \end{cases}$$

and the heat conduction subproblem

$$\begin{cases} \dfrac{\partial(\rho E)}{\partial t} = \dfrac{\partial}{\partial x}(K\dfrac{\partial T}{\partial x}) + \dfrac{2K\frac{\partial T}{\partial x}}{x}, & (4.8a) \\ E = c_v T + \dfrac{1}{2}u^2. & (4.8b) \end{cases}$$

By using uniform space mesh

$$\{x_i \mid x_i = (i - 0.5)h,\ i = 1, 2, \cdots, 6400, h = \frac{5R}{6400}\}$$

and fifth-order FD-WENO scheme, we can easily get the semi-discrete approximation of the PDE subproblem (4.7), and find that it is a mild stiff ODE problem. Similarly, by using second order difference methods on the same uniform space mesh, we can easily get the semi-discrete approximation of the PDE subproblem (4.8), and find that it is a strongly stiff ODE problem. Therefore, the whole semi-discrete approximation of the original PDE problem (4.6) is a composite stiff ODE problem, and we can use CES method to solve it.

However, in order to improve the local accuracy and resolution of the numerical solution of the fluid equation, a more appropriate choice is to use the well known third-order TVD Runge-Kutta method (TVDRK3) to solve the first subproblem, thus we would like to use GCES(TVDRK3,ImEuler) method (rather then CES method) with adaptive time step size to solve this composite stiff ODE problem, and for comparison purposes, we also use the corresponding traditional sequential operator splitting method SOS(TVDRK3,ImEuler) with adaptive time step size to solve the same problem with the same decomposition on the same Dell OptiPlex 755 computer with the same strategy for stopping simplified Newton iteration.

The numerical simulation results obtained by these two different splitting methods at time $t = 2480.03$ are plotted in figures 4.2a and 4.2b, the costed computing time t_{cpu} are listed in table 4.1, from which we see that although the numerical results obtained by these two different splitting

 ISBN: 978-84-608-6082-2

methods are almost the same, but the computational speed of GCES(TVDRK3,ImEuler) method is much faster than that of SOS(TVDRK3,ImEuler) method.

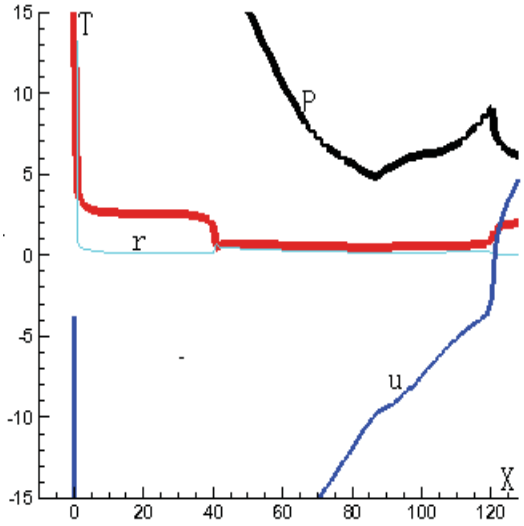


Figure 4.2a Numerical simulation of T, P, ρ, u at time $t = 2480.03$

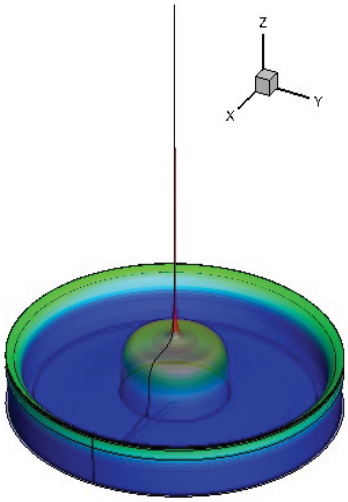


Figure 4.2b Temperature T of the target sphere at time $t = 2480.03$

Table 4.1 Computing time of different methods

Splitting method	computing time t_{cpu}
GCES(TVDRK3,ImEuler)	2 hours 12 minutes 22 seconds
SOS(TVDRK3,ImEuler)	7 hours 42 minutes 12 seconds

Example 4.3 Consider the initial-boundary value problem in nonlinear partial functional differential equations

$$\begin{cases} \dfrac{\partial u}{\partial t} = t^4 \dfrac{\partial}{\partial x}(u\dfrac{\partial u}{\partial x}) - u^2(x,t) - 2\ u^2(x,\dfrac{t}{2}) \\ \quad - (4x^2-4x+1)\ u(x,t-\dfrac{\pi}{2}) + \displaystyle\int_{t-\frac{\pi}{2}}^{t} \cos\theta\cos 2\theta u(x,\theta)d\theta \\ \quad - 16(6x^2-6x+1)t^4\sin^2 t + 16x^2(1-x)^2(1+\sin^2 t), & x\in(0,1),\ t\in[0,\pi], \\ u(0,t) = u(1,t) = 0, & t\in[0,\pi], \\ u(x,t) = 4x(1-x)\sin t, & x\in(0,1),\ t\in[-\frac{\pi}{2},0]. \end{cases} \tag{4.9}$$

This problem has a unique true solution $u(x,t) = 4x(1-x)\sin t$. Using uniform space mesh

$$\{x_i : x_i = ih_x,\ i = 0,1,\cdots,1001,\ h_x = 1/1001\},$$

 ISBN: 978-84-608-6082-2

we get the following semi-discrete problem by difference method

$$\begin{cases} \dfrac{du_i(t)}{dt} = t^4 \dfrac{u_{i+1}^2(t) - 2\ u_i^2(t) + u_{i-1}^2(t)}{2h_x^2} - u_i^2(t) - 2\ u_i^2(\dfrac{t}{2}) \\ \qquad - (4x_i^2 - 4x_i + 1)\ u_i(t - \dfrac{\pi}{2}) + \displaystyle\int_{t-\frac{\pi}{2}}^{t} \cos\theta \cos 2\theta u_i(\theta) d\theta \\ \qquad - 16(6x_i^2 - 6x_i + 1)t^4 \sin^2 t + 16x_i^2(1 - x_i)^2(1 + \sin^2 t), \qquad t \in [0, \pi], \\ u_i(t) = 4x_i(1 - x_i)\sin t, \qquad t \in [-\frac{\pi}{2}, 0], \end{cases} \tag{4.10}$$

where $i = 1, 2, \cdots, 1000$, $u_0(t) = u_{1001}(t) = 0$.

For each time integration step from t_n to t_{n+1}, we firstly decompose the semi-discrete problem (4.10) into two sub-problems, i.e., the non-stiff sub-problemsmall

$$\begin{aligned} \frac{d\bar{u}_i(t)}{dt} &= -\bar{u}_i^2(t) - 2\ \bar{u}_i^2(\frac{t}{2}) \\ &\quad - (4x_i^2 - 4x_i + 1)\ \bar{u}_i(t - \frac{\pi}{2}) + \int_{t-\frac{\pi}{2}}^{t} \cos\theta \cos 2\theta \bar{u}_i(\theta) d\theta \\ &\quad - 16(6x_i^2 - 6x_i + 1)t^4 \sin^2 t + 16x_i^2(1 - x_i)^2(1 + \sin^2 t), \end{aligned} \tag{4.11a}$$

and the strongly stiff sub-problem

$$\frac{d\tilde{u}_i(t)}{dt} = t^4 \frac{\tilde{u}_{i+1}^2(t) - 2\ \tilde{u}_i^2(t) + \tilde{u}_{i-1}^2(t)}{2h_x^2}. \tag{4.11b}$$

Note that for simplicity here we did not write the initial values of the sub-problems. Then we use our CES method with time stepsize $\tau = 10^{-3}, 10^{-4}$, respectively, to solve the semi-discrete VFDE problem (4.10) with decomposition (4.11a)-(4.11b). Furthermore, for comparison purpose, we also use the corresponding traditional sequential operator splitting method SOS(ExEuler,ImEuler) to solve the same semi-discrete problem with the same decomposition and same stepsize on the same Dell OptiPlex 755 computer. The maximum global errors E_{max} of the numerical results on the whole integration interval $[0, \pi]$ with respect to the true solution $u(x, t)$ of the original problem (4.9) and the costed computing time t_{cpu} are listed in Table 4.2, from which we see that for solving nonlinear partial functional differential equation problems of the form (4.9), our CES method is of high speed and can reach the expected high accuracy, whereas the traditional operator splitting method SOS(ExEuler,ImEuler) is failed.

Table 4.2 Maximum global errors E_{max} of the numerical results and computing time t_{cpu} of different splitting methods when applied to problem (4.10)

splitting method	$\tau = 10^{-3}$		$\tau = 10^{-4}$	
	E_{max}	t_{cpu}(sec)	E_{max}	t_{cpu}(sec.)
CES	1.699950×10^{-4}	6.547	1.710165×10^{-5}	52.203
SOS(ExEuler,ImEuler)	floating-point overflow		floating-point overflow	

 ISBN: 978-84-608-6082-2

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No.11171282)

References

[1] A. Araujo, J.R. Branco, J.A. Ferreira, On the stability of a class of splitting methods for integro-differential equations, Applied Numerical Mathematics, 59 (2009), 436–453.

[2] A. Arrarás, L.Portero, J.C.Jorge Locally linearized fractional step methods for nonlinear parabolic problems, J. Comput. Appl. Math., 234(2010), 1117–1128.

[3] U.M.Ascher, S.J.Ruuth, B.T.R.Wetton, Implicit-explicit methods for time-dependent differential equations, SIAM. J. Numer. Anal., 32(1995), 797–823.

[4] A. Batkai, P. Csomos, B. Farkas, Operator splitting for nonautonomous delay equations, Computers and Mathematics with Applications, 65(2013) 315–324.

[5] G.J. Cooper, A. Sayfy, Additive Runge-Kutta methods for stff ordinary differential equations, Matic. Comp., 161(1983), 207–218.

[6] P. Csomos, G. Nickel, Operator splitting for delay equations, Computers and Mathematics with Applications, 55(2008), 2234–2246.

[7] K. Dekker, J.G. Verwer, Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations, North Holland, Amsterdam, 1984.

[8] J. Frank, W. Hundsdorfer, J.G. Verwer, On the stablity of implicit-explicit linear multistep method, Appl. Numer. Math., 25(1997), 193–205.

[9] D. Gasiorowski, Impact of diffusion coefficient averaging on solution accuracy of the 2D nonlinear diffusive wave equation for floodplain inundation, J. Hydrol., 517(2014), 923–935.

[10] J. Geiser, An iterative splitting approach for linear integro-differential equations, Applied Mathematics Letters, 26 (2013), 1048–1052.

[11] J. Geiser, Operator-splitting methods in respect of eigenvalue problems for nonlinear equations and applications for Burgers equations, J. Comput. Appl. Math., 231(2009), 815–827.

[12] J. Geiser, Operator-splitting methods via the Zassenhaus product formula, Appl. Math. Comput., 217(2011), 4557–4575.

[13] H. Guo, J. Zhang, H. Fu, Two splitting positive definite mixed finite element methods for parabolic integro-differential equations, App. Math. Comp., 218(2012), 11255–11268.

[14] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II, Springer-Verlag, Berlin, Heidelberg, 1991.

[15] P. Henrici, Discrete variable methods in ordinary differential equations, John Wiley & Sons, 1962.

[16] L.I. Ignat, A splitting method for the nonlinear Schröinger equation, Sci. J. Differ. Equ., 250(2011), 3022–3046.

[17] K.J. in'tHout, On the contractivity of implicit-explivit linear multistep methods, Appl. Numer. Math., 42(2002), 201–212.

[18] Z. Jackiewicz, H. Liu, B. Li, Y. Kuang, Numerical simulations of traveling wave solutions in a drift paradox inspired diffusive delay population model, Mathematics and Computers in Simulation, 96(2014), 95–103.

 ISBN: 978-84-608-6082-2

[19] O. Koch, Ch. Neuhauser, M. Thalhammer, Embedded exponential operator splitting methods for the time integration of nonlinear evolution equations, Appl. Numer. Math., 63(2013), 14–24.

[20] T. Koto, Stability of IMEX Runge-Kutta methods for delay differential equations, J. Comput. Appl. Math,, 211(2008), 201–212.

[21] J.D. Lambert, Computational Methods in Ordinary Differential Equations, John Wiley & Sons Ltd., 1973.

[22] D. Li, C. Zhang, W. Wang, Y. Zhang, Implicit-explicit predictor-corrector schemes for nonlinear parabolic differential equations, Applied Mathematical Modelling, 35(2011), 2711–2722.

[23] Shoufu Li, Canonical Euler splitting method for nonlinear composite stiff evolution equations, plenary lecture, 12-th National Conference on Numerical Differential Equations, China, 2016. To appear in Appl. Math. Comput.

[24] Shoufu Li, B-theory of Runge-Kutta methods for stiff Volterra functional differential equations, Science in China (Series A), 46(2003), 5:662–674.

[25] Shoufu Li, High Order Contractive Runge-Kutta Methods for Volterra Functional Differential Eqations, SIAM J. Numer. Anal., 47(2010), 6:4290–4325.

[26] Shoufu Li, Classical theory of Runge-Kutta methods for Volterra functional differential equations, Applied Mathematics and Computation, 230(2014), 78-95.

[27] Shoufu Li, B-convergence theory of Runge-Kutta methods for stiff Volterra functional differential equations with infinite integration interval, SIAM J. Numer. Anal., 53(2015), 6: 2570–2583.

[28] Shoufu Li, Generalized canonical Euler splitting methods for composite stiff problems in evolution equations, Report at Huazhong university of science and technology, Wuhan, PR China, 2016. To appear.

[29] J. Makungu, H. Haario, W.C. Mahera, A generalized 1-dimensional particle transport method for convection diffusion reaction modelAfr. Mat., (2012) 23:21–39.

[30] B. Malengier, Parameter estimation in convection dominated nonlinear convection-diffusion problems by the relaxation method and the adjoint equationJ. Comput. Appl. Math., 215(2008), 477–483.

[31] M. Remeškoviá, J.A. Ferreiraa, Solution of convection-diffusion problems with nonequilibrium adsorption, J. Comput. Appl. Math., 169(2004),101–116.

[32] D.R. Reynolds, R. Samtaney, C.S. Woodward, Operator-based preconditioning of stiff hyperbolic systems, SIAM J. Sci. Comp., 32(2010), 1:150–170.

[33] L.F. Shampine, C.W. Gear, A user's view of solving stiff ordinary differential equations, SIAM Rev., 21(1979) 1–17.

[34] A. Singh, R.M. Allen-King, A.J. Rabideau, Groundwater transport modeling with nonlinear sorption and intraparticle diffusion, Sci. Adv. Water Resour., 70(2014), 12–23.

[35] Y.N. Skiba, D.M. Filatov, Splitting-based schemes for numerical solution of nonlinear diffusion equations on a sphere, Appl. Math. Comput., 219(2013), 8467–8485.

[36] G. Strang, On the construction and comparison of difference schemes, SIAM J. Numer. Anal. 5(1968), 506-517.

[37] J.G. Verwer, B.P. Sommeijer, W. Hundsdorfer, RKC time-stepping for advection-diffusion-reaction problems, Sci. JCP, 201(2004), 61–79.

[38] M. Založnik, H. Combeau, An operator splitting scheme for coupling macroscopic transport and grain growth in a two-phase multiscale solidification model: Part I-Model and solution scheme, Sci. CMS., 48(2010), 1–10.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Accuracy analysis of a 2D adaptive mesh refinement method using lid-driven cavity flow and two refinements

Zhenquan Li[1] **and Robert Wood**[1]

[1] *School of Computing and Mathematics, Charles Sturt University*

emails: jali@csu.edu.au, rwood@csu.edu.au

Abstract

Locating accurate centres of vortices is one of the accurate measures for computational methods in fluid flow and the lid-driven cavity flows are widely used as benchmarks. This paper analyses the accuracy of an adaptive mesh refinement method using 2D steady incompressible lid-driven cavity flows for two refinements. The adaptive mesh refinement method performs mesh refinement based on the numerical solutions of Navier-Stokes equations solved by `Navier2D`, a vertex centred Finite Volume that uses the median dual mesh to form the Control Volumes (CVs) about each vertex. The accuracy of the refined meshes is demonstrated by the centres of vortices obtained in the benchmarks being contained in the twice refined grids. The adaptive mesh refinement method investigated in this paper is proposed based on the qualitative theory of differential equations. Theoretically infinite refinements can be performed on an initial mesh. Practically we can stop the process of refinement based on tolerance conditions. The method can be applied to find the accurate numerical solutions of any mathematical models containing continuity equations for incompressible fluid, steady state fluid flows or mass and heat transfer.

Key words: adaptive mesh refinement, finite volume method, lid-driven cavity flow

1 Introduction

Meshing is the process of breaking up a physical domain into finite smaller sub-domains (called elements, cells or grids) in order to evaluate the discrete numerical solutions of differential equations at the nodes. Adaptive mesh refinement is a computational technique to improve the accuracy of the numerical solutions by starting the calculations on a coarse initial mesh and then refining this mesh based on refinement criteria.

  ISBN: 978-84-608-6082-2

There are a large number of publications on adaptive mesh refinements and their applications. Some refinement methods use a refinement criterion which is based on local truncation errors (e.g. Almgren, Bell, Colella, Howell & Welcome [1]). Other common methods include the so-called h-refinement (e.g. Lohner [19]), p-refinement (e.g. Bell, Berger, Saltzman & Welcome [3]) or r-refinement (e.g. Miller & Miller [20]), with different combinations of these also possible (e.g. Capon & Jimack [4]). The overall aim of these adaptive algorithms is to allow a balance to be obtained between accuracy and computational efficiency in solving differential equations.

We proposed adaptive mesh refinement methods for 2D velocity fields (Li [15]) and for 3D fields (Li [14]) based on a theorem in qualitative theory of differential equations (Theorem 1.14, page 18, Ye et al. [22]). The theorem indicates that a divergence free vector field has no limit cycles or one sided limit cycles, that is, the trajectories (or streamlines) of divergence free vector fields are closed curves in bounded domains (singular points are streamlines) that have also been shown by benchmarks (e.g. Erturk et al. [6]). The adaptive mesh refinement methods adaptively refine meshes based on the information of evaluated numerical velocity fields to obtain refined meshes on which the linear interpolation of the numerical velocity fields approximates continuous divergence free vector fields. The area on which the linear interpolation is not equivalent to a divergence free vector field reducibly closes to zero when the number of refinements increases.

Identification of accurate locations of singular points and asymptotic lines (planes), and drawing closed streamlines are some of the accuracy measures for computational methods. Using numerical velocity fields obtained by taking the vectors of the analytical velocity fields at nodes of the refined meshes, examples show the accuracy of the adaptive mesh refinement methods include: locating the singular points and asymptotic lines for 2D [12]; the singular points and asymptotic plane for 3D [13]; and drawing closed streamlines (Li [12], [13]). We showed that the once refined mesh for 2D velocity fields provides accurate estimates for the singular points of 2D steady incompressible lid-driven cavity flows using the numerical velocity fields (Lal & Li [10]). The numerical velocity fields are obtained by solving the Navier-Stokes equations with the boundary conditions numerically using a second order colocated finite volume method (GSFV) with a splitting method for time discretization (Faure, Laminie & Temam [7]). We applied the adaptive mesh refinement method to the initial meshes and the numerical velocity fields, and take the centres of refined grids in the vortex regions as the estimates of the singular points. The comparison of the estimates with the benchmarks shows that the estimates for the singular points are accurate.

Mesh refinement is necessary for producing accurate numerical solutions. Li [16] considers 2D lid-driven cavity flows using finer meshes 99×99 for $\mathrm{Re} = 1000$, 121×121 for $\mathrm{Re} = 2500$ and 139×139 for $\mathrm{Re} = 5000$. The results show that the different sizes of vortices (primary, secondary, tertiary and quaternary vortices) require different densities of mesh nodes in the separated-flow regions for similar relative errors of centre locations. The same

conclusion is derived in Armaly et al. [2]. An investigation starting from coarser initial meshes and demonstrating the centres of vortices are contained in the refined grids of once refined meshes has been done (Li et al. [18]).

This paper reports the accuracy analysis of the same adaptive mesh refinement method for 2D proposed by Li [15] for more refinements using the benchmarks for 2D lid-driven cavity flows. We show that the centres of vortices obtained in the benchmarks are contained in the twice refined grids for Re = 100 and Re = 1000. We conclude that more accurate centres of vortices can be achieved when more refinements are performed.

2 Review of algorithm of adaptive mesh refinement

This section summarizes the adaptive mesh refinement method proposed by Li [15] based on Theorem 1.14 of [22].

Assume that $\mathbf{V}_l = \mathbf{AX} + \mathbf{B}$ is a vector field obtained by linearly interpolating the vectors at the three vertexes of a triangle, where

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

is a matrix of constants,

$$\mathbf{B} = \begin{pmatrix} b_1' \\ b_2' \end{pmatrix}$$

is a vector of constants, and $\mathbf{X} = (x_1, x_2)^T$. The vector $\mathbf{V}_l$ is unique if the area of the triangle is not zero [11]. The continuity equation for $\mathbf{V}_l$ and a steady flow or an incompressible fluid is

$$\nabla \cdot \mathbf{V}_l = \text{trace}(\mathbf{A}) = 0. \tag{1}$$

Let f be a scalar function depending only on spatial variables. We assume that $f\mathbf{V}_l$ is divergence free and then calculate the expressions of f. Li [15] derives the expressions of f for the four different Jacobian forms of the coefficient matrix $\mathbf{A}$ as shown in Table 1. Variables y_1 and y_2 in Table 1 are the components of $(y_1, y_2)^T = \mathbf{V^{-1}X}$ where $\mathbf{V}$ satisfies $\mathbf{AV} = \mathbf{VJ}$ and $\mathbf{J}$ is one of the Jacobian matrices in Table 1. Vectors $\mathbf{V}_l$ and $f\mathbf{V}_l$ produce same streamlines if $f \neq 0, \infty$ (refer to Section 2.2 of [12]). Therefore, if $f\mathbf{V}_l$ is divergence free, $\mathbf{V}_l$ produces divergence free streamlines. The functions f are calculated by solving differential equations [17]. Scalar functions f reduce the number of refined grids in refined meshes [17]. The conditions (MC)(MC is the abbreviation of mass conservation) are the functions f in Table 1 not equaling zero or infinity at any point on the triangular domains when $f\mathbf{V}_l$ is divergence free on these triangular domains.

We review the algorithm of adaptive mesh refinement for quadrilateral meshes [18]. The algorithm is also applicable to a triangular mesh after a subdivision of a triangle to

Table 1: Jacobian matrices and corresponding expressions of f ($C \neq 0$)

Case	Jacobean	f
1	$\begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix}$ $(0 \neq r_1 \neq r_2 \neq 0)$	$\frac{C}{\left(y_1+\frac{b_1}{r_1}\right)\left(y_2+\frac{b_2}{r_2}\right)}$
2	$\begin{pmatrix} r_1 & 0 \\ 0 & 0 \end{pmatrix}$ $(r_1 \neq 0)$	$\frac{C}{y_1+\frac{b_1}{r_1}}$
3	$\begin{pmatrix} r_1 & 0 \\ 0 & r_1 \end{pmatrix}$ $(r_1 \neq 0)$	$\frac{C}{\left(y_1+\frac{b_1}{r_1}\right)^2}$
4	$\begin{pmatrix} \mu & \lambda \\ -\lambda & \mu \end{pmatrix}$ $(\mu \neq 0, \lambda \neq 0)$	$\frac{C}{\left(y_1+\frac{\mu b_1-\lambda b_2}{\mu^2+\lambda^2}\right)^2+\left(y_2+\frac{\lambda b_1+\mu b_2}{\mu^2+\lambda^2}\right)^2}$

a number of small triangles is defined. The following grid refinement algorithm describes how to use the conditions (MC) to refine a quadrilateral grid in a given mesh. To avoid an infinite refinement of the mesh, we choose a pre-specified threshold number of refinements T based on the accuracy requirements. The algorithm of grid refinement is:

Step 1 Subdivide a quadrilateral grid into two triangles. If $\mathbf{V}_l$ satisfies Eq. (1) on both triangles, no refinement for the grid is required. Otherwise, go to Step 2;

Step 2 Apply the conditions (MC) to both of the triangles. If the conditions (MC) are satisfied on both triangles, no refinement for the grid is required. Otherwise, we subdivide the grid into a number of small grids such that the lengths of all sides of the small grids are truly reduced (e.g. connecting the mid-points of opposite sides of a quadrilateral by line segments produces four small quadrilaterals and the lengths of the sides of the four small quadrilaterals are truly reduced).

In this paper, we subdivide a quadrilateral grid by connecting the mid-points of two opposite sides of a quadrilateral.

The algorithm of adaptive mesh refinement is:

Step 1 Evaluate the numerical velocity field for a given initial mesh;

Step 2 Refine all grids of the initial mesh one by one using the above algorithm of grid refinement;

Step 3 Take the refined mesh as initial mesh and go to Step 1 until a satisfactory numerical velocity field is obtained or the threshold number T is reached.

The abbreviations BR, BL and TL refer to bottom right, bottom left and top left corners of the cavity, respectively. The number following these abbreviations refer to the vortices that appear in the flow, which are numbered according to size (for example, BR1 refers to bottom right secondary vortex).

3 Accuracy analysis by comparisons with benchmarks

In this section, we use `Navier2D`, an open source MATLAB CFD Codes by Darren Engwirda for evaluating the numerical velocity field on triangular meshes [5]. The results reported in this paper have the residuals for both x and y being less than 10^{-6} in the evaluation of the numerical velocity fields.

One of the possible comparisons is the adaptive mesh refinement which enforces refinement criteria

$$\|\nabla u^{(k)}\| \leq \epsilon \|u^h\|_1$$

everywhere in the mesh, where $\|\cdot\|$ is the L_2 norm, $\|\cdot\|_1$ is the H^1 norm, ϵ is the discretization tolerance, u^h is finite-dimensional approximation for u, and k in $\|\nabla u^{(k)}\|$ is the number of subdomains (Henderson [9], 293–299). Even though there might be some relations between the refined meshes and the vorticity field as ϵ decreases, no information is provided on the pattern of the flow field such as locations of the centres of vortices [18].

We take the case for Re = 100 as an example to show how we switch between triangular meshes to quadrilateral meshes. Fig. 1 is the initial mesh with size 25×25 uniform grids. The triangulated initial mesh is obtained by connecting bottom left vertex to top right vertex by a line segment in each grid of the initial mesh and then is loaded it into `Navier2D` for the first evaluation of the numerical velocity field.

Fig. 2 shows the once refined mesh using the evaluated numerical velocity field at the nodes. Fig. 3 shows the triangulated mesh from the once refined mesh shown in Fig. 2. The triangulated mesh is loaded into `Navier2D` for the second evaluation of the numerical velocity field.

We take the results for Re = 100 [8, 21], and the results for Re = 1000 [6] as the benchmarks for accuracy analysis. We consider the accuracy of the adaptive mesh refinement method using the inclusion of the centres of vortices identified in the benchmarks in refined grids.

3.0.1 Re = 100

Fig. 4 shows the twice refined mesh using the second evaluated numerical velocity field at the nodes shown in Fig. 3 and the centres of vortices (black dots) given by the benchmarks [8]. All three centres (primary vortex, BL1 and BR1) are contained in the twice refined grids of the two refined mesh.

3.0.2 Re = 1000

This section shows the figures for Re = 1000 generated from an initial mesh with 35×35 uniform grids. Fig. 5 shows the twice refined mesh and the centres of vortices (black dots)

 ISBN: 978-84-608-6082-2

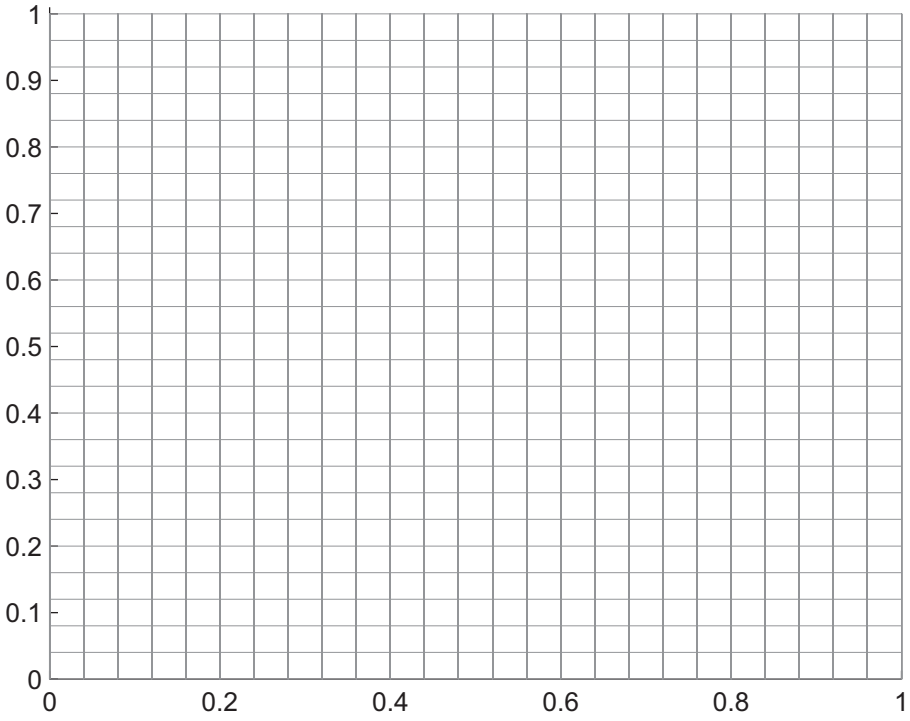


Figure 1: Initial mesh for Re = 100 with size 25×25.

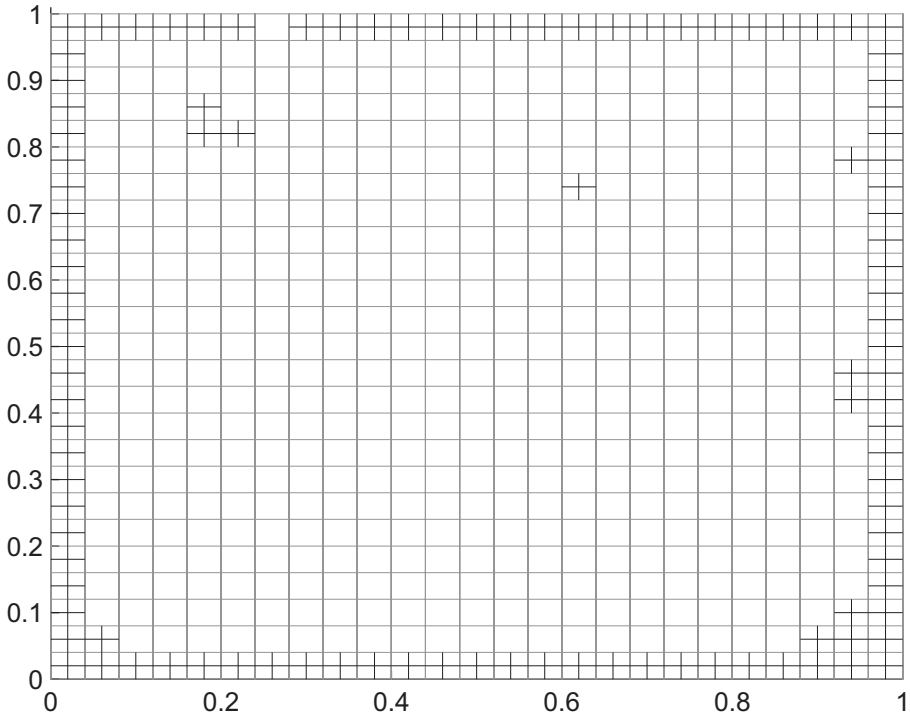


Figure 2: Once refined mesh for Re = 100.

 ISBN: 978-84-608-6082-2

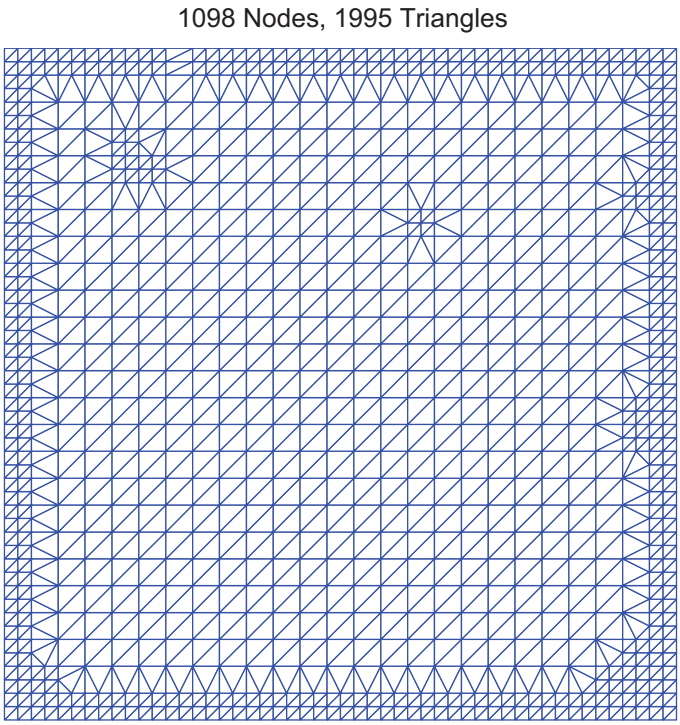


Figure 3: Triangulated mesh based on the mesh shown in Fig. 2

 ISBN: 978-84-608-6082-2

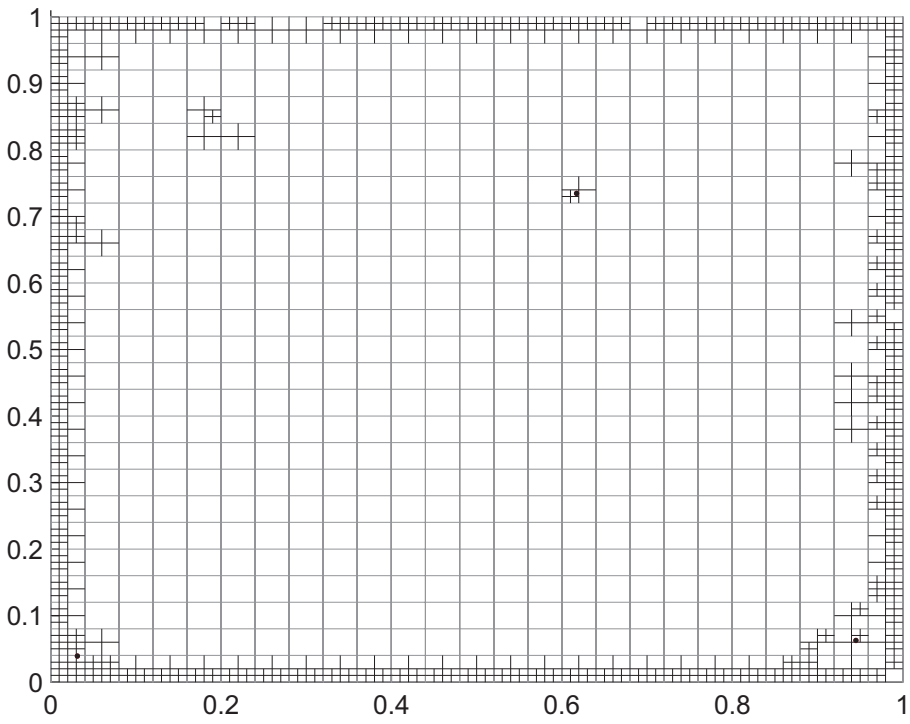


Figure 4: Twice refined mesh for Re = 100 with initial mesh size 25×25.

 ISBN: 978-84-608-6082-2

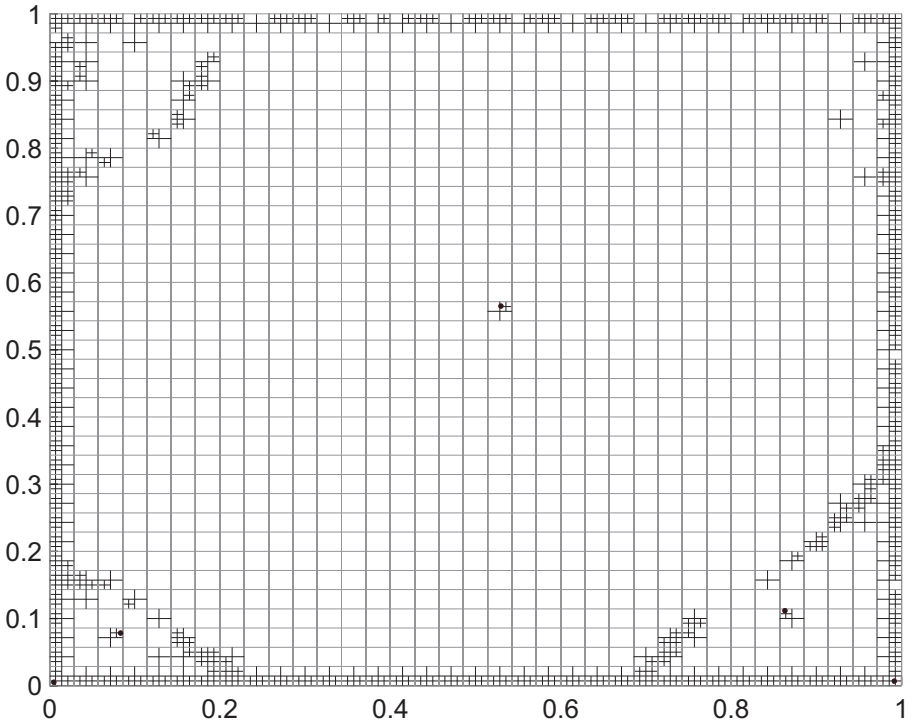


Figure 5: Twice refined mesh for Re = 1000 with initial mesh size 35×35.

given by the benchmarks [6]. All five centres (primary vortex, BL1, BR1, BL2 and BR2) are contained in the twice refined grids.

4 Conclusion

We applied the adaptive mesh refinement method twice to the initial meshes based on the information of numerical solutions of 2D lid-driven cavity flows using `Navier2D`. We demonstrate the accuracy of the adaptive mesh refinement method by the inclusion of the centres of vortices in twice refined grids of twice refined meshes. If we refine the initial meshes more times, we obtain more accurate estimates for the centres of vortices. We are able to achieve the required accuracy for the centres of vortices by selecting an appropriate threshold number T.

References

[1] A. Almgren, J. Bell, P. Colella, L. Howell, M.A. Welcome), *Conservative adaptive projection method for the variable density incompressible Navier–Stokes equations*, J. Comput. Phys. **142** (1998) 1–46.

[2] B.F. Armaly, F. Durst, J.C.F. Pereir, B. Schönung, *Experimental and theoretical investigation of backward-facing step flow*, J. Fluid Mech. **127** (1983) 473–496.

[3] J. Bell, M. Berger, J. Saltzman, M. Welcome, *Three dimensional adaptive mesh refinement for hyperbolic conservation laws*, J. of Sci. Comput. **15** (1994) 127–138.

[4] P.J. Capon, P.K. Jimack, *An adaptive finite element method for the compressible Navier–Stokes equations*, In: Numerical Methods for Fluid Dynamics, Baines MJ, Morton KW (eds): OUP, 1995; 327–333.

[5] D. Engwirda, *Navier-Stokes solver (Navier2d)*, MATLAB Central File Exchange, 2006.

[6] E. Erturk, T.C. Corke, G. Gökcöl, *Numerical solutions of 2-D steady incompressible driven cavity flow at high Reynolds numbers*, Int. J. Numer. Meth. Fl., **48** (2005) 747–774.

[7] S. Faure, J. Laminie, R. Temam, *Colocated finite volume schemes for fluid flows*, Commun. Comput. Phys. **4** (2008) 1–25.

[8] U. Ghia, K.N. Ghia, C.T. Shin, *High-Re solutions for incompressible flow using the Navier-Stokes equations in vorticity-velocity variables*, J. Comput. Phys. **48** (1982) 387–411.

[9] R.D. Henderson, *Adaptive spectral element methods for turbulence and transition*, In: High-Order Methods for Computational Physics, Barth TJ, Deconinck H (eds): Springer-Verlag Berlin, 1999; 225–324.

[10] R. Lal, Z. Li, *Sensitivity analysis of a mesh refinement method using the numerical solutions of 2-D steady incompressible driven cavity flow*, J. Math. Chem. **53** (2015 844–867.

[11] Z. Li, *A mass conservative streamline tracking method for two-dimensional CFD velocity fields*, J. Flow Visual. Image Process. **9** (2002) 75–87.

[12] Z. Li, *An adaptive streamline tracking method for two-dimensional CFD velocity fields based on the law of mass conservation*, J. Flow Visual. Image Process. **13** (2006) 1–14.

 ISBN: 978-84-608-6082-2

[13] Z. Li, *An adaptive streamline tracking method for three-dimensional CFD velocity fields based on the law of mass conservation*, J. Flow Visual. Image Process. **13** (2006) 359–376.

[14] Z. Li, *An adaptive three-dimensional mesh refinement method based on the law of mass conservation*, J. Flow Visual. Image Process. **14** (2007) 375–395.

[15] Z. Li, *An adaptive two-dimensional mesh refinement method based on the law of mass conservation*, J. Flow Visual. Image Process. **15** (2008) 17–33.

[16] Z. Li, *Accuracy analysis of a mesh refinement method using benchmarks of 2-D lid-driven cavity flows and finer meshes*, J. Math. Chem. **52** (2014) 1156–1170.

[17] Z. Li, G. Mallinson, *Simplification of an existing mass conservative streamline tracking method for two-dimensional CFD velocity fields*, In: GIS and Remote Sensing in Hydrology, Water Resources and Environment, Chen Y, Takara K, Cluckies ID, De Smedt FH (eds), IAHS Press, 2004; 269–275.

[18] Z. Li, R. Wood, *Accuracy analysis of an adaptive mesh refinement method using benchmarks of 2-D steady incompressible lid-driven cavity flows and coarser meshes*, J. Comput. Appl. Math. **275** (2015) 262–271.

[19] R. Lohner, *An adaptive finite element scheme for transient problems in CFD*, Comput. Method Appl. M. **61** (1987) 323–338.

[20] K. Miller, R. Miller, *Moving finite elements, Part I.*, SIAM J. Numer. Anal. **18** (1981) 1019–1032.

[21] M. Sahin, R.G. Owens, *A novel fully implicit finite volume method applied to the lid-driven cavity problem - Part I: High Reynolds number flow calculations*, Int. J. Numer. Meth. Fluids **42** (2003) 57–77.

[22] Y. Ye, Others, *Theory of limit cycles*, American Mathematical Society Press, 1986.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Nodal surfaces in quasi-exactly solvable models

Pierre-François Loos[1]

[1] *Research School of Chemistry, Australian National University, Canberra ACT 2601, Australia*

emails: pf.loos@anu.edu.au

Abstract

We study the nodes of two-, three- and four-electron systems in various ferromagnetic configurations (sp, p^2, sd, pd, p^3, sp^2 and sp^3). In some particular cases (sp, p^2, sd, pd and p^3), we rigorously prove that the non-interacting wave function has the same nodes as the exact (yet unknown) wave function. The number of atomic and molecular systems for which the exact nodes are known analytically is very limited and we show here that this peculiar feature can be attributed to interdimensional degeneracies.

Key words: Fermionic nodes, interdimensional degeneracy, quasi-exactly solvable model

1 Background

Diffusion Monte Carlo (DMC) is a stochastic projector technique [3, 13], and the fixed-node (FN) error is the only systematic error in DMC resulting from the use of a trial wave function Ψ_T with approximate nodes (or zeros). If Ψ_T has the correct nodes, FN-DMC yields the exact energy with only a statistical error that can be made arbitrarily small. Because the FN error is only proportional to the square of the nodal displacement error [2] in a region of the configurational space where the wave function is small, at first glance, it could be thought to have marginal impact. However, it is not true in practice [12, 11, 4]. The fundamental reason is that the positions of the zeroes of the wave function indirectly determine the location of the maxima (in analogy with the vibrating string). Except in some particular cases, electronic or more generally fermionic nodes are poorly understood due to their high dimensionality and complex topology [2, 1, 5]. In this talk, we will study the topology of the nodes in a class of systems composed of same-spin electrons located on the surface of a sphere, as system known to be quasi-solvable [7, 8, 10].

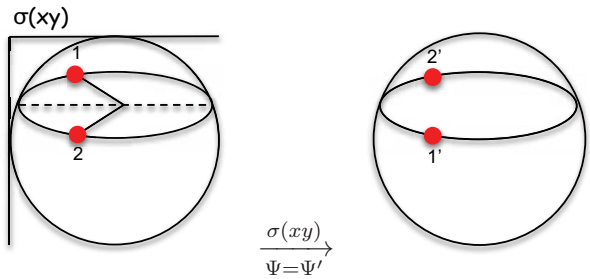


Figure 1: Proof of the exactness of the non-interacting nodes of the ${}^3P^{\mathrm{o}}(sp)$ state.

2 Electrons on a sphere

The model consists of n spin-up electrons restricted to remain a surface of a sphere of unit radius [6, 9]. The non-interacting orbitals for an electron on a unit sphere are the normalized spherical harmonics $Y_{\ell m}(\mathbf{\Omega})$, where $\mathbf{\Omega} = (\theta, \phi)$ are the polar and azimuthal angles respectively. We will label spherical harmonics with $\ell = 0, 1, 2, 3, 4, \ldots$ as s, p, d, f, g, ... functions. The coordinates of the electrons on the sphere of unit radius are given by their cartesian coordinates $x = \cos\phi \sin\theta$, $y = \sin\phi \sin\theta$ and $z = \cos\theta$. Here, I propose to study the nodes of the non-interacting wave function Ψ_0. Because we only consider ferromagnetic systems, Ψ_0 is a Slater determinant of spin orbitals.

We will label each state using the following notations: ${}^{2\mathcal{S}+1}L^{\mathrm{e,o}}$, where $L = S, P, D, F, \ldots$ and $\mathcal{S} = \sum_{i=1}^{n} s_i$ is the total spin angular momentum. The suffixes e (even) and o (odd) are related to the parity of the states given by $(-1)^{\ell_1 + \cdots + \ell_n}$.

As an example, we study the triplet state ${}^3P^{\mathrm{o}}(sp)$, which has the following non-interacting wave function:

$$\Psi_0(sp) = \begin{vmatrix} 1 & z_1 \\ 1 & z_2 \end{vmatrix} = \boldsymbol{z} \cdot \boldsymbol{r}_{12}, \tag{1}$$

where $\boldsymbol{z} = (0, 0, 1)$ is the unit vector of the z axis and $\boldsymbol{r}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$. Due to its ferromagnetic nature, this state has "Pauli" nodes which corresponds to configurations where two electrons touch. The Pauli hyperplanes are only a subset of the full nodes.

Equation (1) shows that the non-interacting nodes of the sp configuration corresponds to small circle perpendicular to the z axis. Now, let us prove that these non-interacting and exact nodes are identical. We begin by placing the two electrons on a small circle perpendicular to the z axis, as sketched in Fig. 1. For this particular configuration, the two electrons have the same value of the polar angle $\theta = \theta_1 = \theta_2$ and, without loss of generality, the azimuthal angles can be chosen such that $\phi_1 = -\phi_2 = \phi$. Suppose that for this configuration the exact wave function has a value $\Psi \equiv \Psi(\{(\theta, +\phi), (\theta, -\phi)\}) = K$. Now, we reflect the wave function with respect to the symmetry plane $\sigma(xz)$ that passes through

 ISBN: 978-84-608-6082-2

the x and z axes and bisects the azimuthal angle ϕ. Due to the P nature of the state, the wave function is invariant to such reflexion, i.e. $\Psi' \equiv \Psi(\{(\theta,-\phi),(\theta,+\phi)\}) = K$. However, the two electrons have been exchanged and because this is a triplet state, the wave function must have changed sign (Pauli principle). Because $\Psi = -\Psi'$, this implies that $K = -K$ which means that $K = 0$ and $\forall(\theta,\phi), \Psi(\theta,\theta,\phi,-\phi) = 0$. We have just discovered the nodes of the *sp* configuration by using simple symmetry operations!

Acknowledgements

The author would like to thank the Australian Research Council for a Discovery Early Career Researcher Award (DE130101441) and a Discovery Project grant (DP140104071).

References

[1] Bajdich, M., Mitas, L., Drobny, G., and Wagner, L. K. *Approximate and exact nodes of fermionic wavefunctions: Coordinate transformations and topologies,* Phys. Rev. B **72** (2005) 075131.

[2] Ceperley, D. M. *Fermion nodes,* J. Stat. Phys. **63** (1991) 1237–1267.

[3] Ceperley, D. M., and Kalos, M. H. *Monte Carlo Methods in Statistical Physics.* Springer Verlag, Berlin, 1979.

[4] Kulahlioglu, A. H., Rasch, K. M., Hu, S., and Mitas, L. *Density dependence of fixed-node errors in diffusion quantum monte carlo: Triplet pair correlations,* Chem. Phys. Lett. **591** (2014) 170–174.

[5] Loos, P. F., and Bressanini, D. *Nodal surfaces and interdimensional degeneracies,* J. Chem. Phys. **142** (2015) 214112.

[6] Loos, P. F., and Gill, P. M. W. *Ground state of two electrons on a sphere,* Phys. Rev. A **79** (2009) 062517.

[7] Loos, P. F., and Gill, P. M. W. *Two electrons on a hypersphere: A quasiexactly solvable model,* Phys. Rev. Lett. **103** (2009) 123008.

[8] Loos, P. F., and Gill, P. M. W. *Excited states of spherium,* Mol. Phys. **108** (2010) 2527–2532.

[9] Loos, P. F., and Gill, P. M. W. *Thinking outside the box: The uniform electron gas on a hypersphere,* J. Chem. Phys. **135** (2011) 214111.

 ISBN: 978-84-608-6082-2

[10] Loos, P. F., and Gill, P. M. W. *Exact wave functions of two-electron quantum rings,* Phys. Rev. Lett. **108** (2012) 083002.

[11] Rasch, K. M., Hu, S., and Mitas, L. *Fixed-node errors in quantum monte carlo: Interplay of electron density and node nonlinearities,* J. Chem. Phys. **140** (2014) 041102.

[12] Rasch, K. M., and Mitas, L. *Impact of electron density on the fixed-node errors in quantum monte carlo of atomic systems,* Chem. Phys. Lett. **528** (2012) 59–62.

[13] Reynolds, P. J., Ceperley, D. M., Alder, B. J., and Lester, Jr., W. A. *Fixed-node quantum monte carlo for molecules,* J. Chem. Phys. **77** (1982) 5593–5603.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

How group size influences the efficiency of FMM

J. A. López-Fernández[1], M. López-Portugués[1], José Ranilla[2], R. González-Ayestarán[1] and F. Las-Heras[1]

[1] *Departamento de Ingeniería Eléctrica, Electrónica, de Computadores y Sistemas, Universidad de Oviedo*

[2] *Departamento de Informática, Universidad de Oviedo*

emails: jelofer@uniovi.es, lopezpmiguel@uniovi.es, ranilla@uniovi.es, rayestaran@uniovi.es, flasheras@uniovi.es

Abstract

The Fast Multipole Method (FMM) is commonly used to speed-up the solution of a variety of N-body type problems. The FMM needs to divide the geometry of interest into groups and the size of these groups (D), unknown beforehand, may have a deep impact on the time to solution. Nonetheless, octree structures may be used during the setup of the solver to divide the geometry, into groups of size D, and to estimate the cost of the FMM. In this paper, we use octree structures to efficiently look for the size D that minimizes the time cost of two parallel frameworks of the FMM: single level FMM and Fast Fourier Transform FMM (FMM-FFT). We found that the reduction of the time cost is highly appreciated when the adequate D is considered.

Key words: FMM, FMM-FFT, cube size, computational cost

1 Introduction

Since the early 1990s, the Fast Multipole Method (FMM) has been widely used to speed-up the solution of a variety of N-body type problems [1, 2]. In fact, it is possible to find FMM approaches of those problems in different engineering fields such as: electromagnetics [2, 3, 4, 5, 6], acoustics [7, 8], or mechanics [9]. In the aforementioned problems the geometry of interest is discretized into a number, N, of elements (or basis functions) whose interactions may be computed by means of a Matrix-Vector Product (MVP). In fact, iterative solutions of the above problems usually require at least one MVP per iteration. The FMM avoids

 ISBN: 978-84-608-6082-2

the explicit calculation of each MVP by sparsifying the involved matrix [10]. This efficient computation of each MVP relies on a fast evaluation of the interactions among clusters of elements that are far from each other. The elements are clustered taking into consideration the spatial proximity between them [10]. The use of octree structures [11] yields a very efficient clustering or grouping [12].

During these years of FMM usage, a number of different versions of the algorithm have been developed, among others: single level FMM [2], Multi-Level FMM (MLFMM, see [13] and [3]), Fast Fourier Transform FMM (FMM-FFT, see [14]), and *Nested* FMM-FFT [15]. From the computational point of view, more specifically from the parallel programming, each of the related versions of the FMM has its own pros and cons. On the one hand, the single level FMM has a bigger time complexity than the others although it is the more prone to parallelize and seems to be very adaptive to different hardware architectures [12, 8]. On the other hand, the MLFMM is the most efficient in terms of time complexity, although it shows a poor scalability on distributed memory systems [4]. Meanwhile, the FMM-FFT and the *Nested* FMM-FFT, reduce the computational complexity of the single level FMM without dramatically worsening the parallelization properties of the latter.

To achieve the theoretical computational complexity of the FMM, the number of groups, N_g, must be chosen carefully. For instance, in the case of the single level FMM it is recommended to use $N_g \propto \sqrt{N}$ [10]. In the case of MLFMM, the edge length of the cube at the finest level is about one half of a wavelength [3] and, therefore, it contains just a few elements. Although these general recommendations are fine to achieve the asymptotic cost of the algorithm, they may not be adequate when the goal is to achieve the smallest time to solution for a given problem. In fact, we have observed that the group size that produces the most efficient performance of the FMM depends on several factors, such are: the version of the algorithm, the size of the problem, and some ground-level details of its implementation.

Nonetheless, to the best of our knowledge there is not a published research on a recommended procedure to get the optimal group size. In fact, some of the papers published in this matter do not even specify the group size used in the numerical experiments, or they use a fix group size (no matter which variant of the FMM or which architecture has been used to solve the problem) or they use a variable group size but without specifying the criteria employed to choose the size. The use of octree structures [11] make possible to obtain a very accurate and fast estimation of the time and memory required by the FMM solver for a given problem and cube edge length (group size), D.

In Section 2 of this work we briefly describe our procedure to efficiently estimate de group size that minimizes the time cost. Section 3 includes some results that show the convenience of considering a group size adapted to the implementation and to the problem size. Finally, Section 4 summarizes some conclusions of the presented work.

2 Selection of the optimal cube size

For the sake of illustration, we apply our technique of optimal cube size selection to an acoustic scattering problem. The goal is to find the acoustic pressure on the surface of an obstacle that is impinged by an acoustic wave. We discretize the Burton and Miller equation [16] to formulate the problem as a linear system of N equations with N unknowns. We use the Generalized Minimum RESidual (GMRES, see [17]) to iteratively solve the aforementioned system of equations. Each GMRES iteration requires a MVP that is speed-up by means of the FMM (or the FMM-FFT). Among the different parts on which the algorithm may be divided, the MVPs consume the highest proportion of the time to solution. As a consequence, we focused on the minimization of the MVP (or iteration) time cost. In order to estimate the single level FMM time cost, we have used the model described in Section 4 of a previous work of the authors [12]. The estimation of the FMM-FFT time cost is performed in a similar fashion.

The developed procedure to choose the optimal group size is divided into two different steps: one initial estimation and one fine estimation based on a guided search.

The first step is based on an estimation of the number of non-empty groups (N_g) and the number of neighboring groups, so any octree generation is avoided. Since this initial estimation may be inaccurate, it is only used to demarcate the limits of the guided search of the second step.

In the second step, and taking into account the limits obtained from the first step, different cube sizes are tested. Now, the octree related to each cube size is generated in order to precisely calculate the MVP time (that may be derived from the number of operations required by each step of the FMM or FMM-FFT). First, a coarse search, where the cube size is increased by 0.5λ at a time, is carried out. Then, using the best cube size obtained from the coarse search as a a central point, a fine search, where the cube size is increased by 0.01λ at a time, is performed over a search domain of 1λ around the central point.

3 Results

The results shown in this work have been obtained using a workstation with one Intel Core i7-3820[1] CPU and 64 GB of RAM. Both FMM and FMM-FFT follow a hybrid parallel design and Intel icc 12.1 and MPICH2 1.5 were used for generating the binaries. In addition, the FFTW 3.3.1 library was used in the FMM-FFT implementation. It is also worth mentioning that single-precision arithmetic has been used in all the codes.

Below, several results for both FMM and FMM-FFT implementations are discussed. In order to show the accuracy of the optimal cube size selection algorithm, both estimated ($\widetilde{t}_{\mathrm{it}}$)

[1] 4 cores at 3.6 GHz (Hyper-Threading and Turbo Boost enabled).

 ISBN: 978-84-608-6082-2

and measured (t_{it}) execution times per iteration are shown. In addition, different group sizes (D/λ) —resulting in different number of groups N_g— are used to show the importance of choosing the appropriate cube size depending on the algorithm and the problem to solve. Four group sizes have been chosen as follows: one small group size ($D/\lambda = 0.5$), the optimal group size for the FMM-FFT implementation (it varies according to the problem), the optimal group size for the FMM implementation (it also varies according to the problem and it is usually bigger than the FMM-FFT one), and one big group size ($D/\lambda = 2.5$).

Table 1: FMM and FMM-FFT parameters and execution times. ⌀2 m sphere analyzed at 4 kHz and discretized using 64272 unknowns.

$N = 64272$

	FMM			FMM-FFT		
D/λ	N_g	$\widetilde{t}_{\text{it}}$	t_{it}	N_g	$\widetilde{t}_{\text{it}}$	t_{it}
0.5	9274	5.75 s	6.51 s	9274	1.92 s	2.04 s
0.62	-	-	-	6122	1.60 s	1.76 s
0.91	2940	2.74 s	2.60 s	-	-	-
2.5	392	11.57 s	11.54 s	392	11.57 s	11.51 s

Table 1 shows the parameters and execution times of a ⌀2 m sphere analyzed at 4 kHz and discretized using 64272 unknowns. Results for both FMM and FMM-FFT implementations are shown. In this problem, the times estimated at runtime are accurate for the different implementations and groups sizes. The cost of the optimal cube size selection is 0.23 s for the FMM and 0.38 s for the FMM-FFT. The FMM-FFT implementation is faster than the FMM one for all the analyzed group sizes, especially when small groups are used. It is also worth noting that the execution time of the FMM with optimal cube size selection ($D/\lambda = 0.91$) is comparable to that of the FMM-FFT when using a small but suboptimal cube size ($D/\lambda = 0.5$).

In Table 2, the configuration parameters and execution times of a ⌀2 m sphere analyzed at 8 kHz and discretized using 256020 unknowns are shown. The cost of the optimal cube size selection is 0.82 s for the FMM and 0.86 s for the FMM-FFT. In this problem, the time estimated at runtime for the FMM is not much accurate when the cube size is small ($D/\lambda = 0.5$). Nevertheless, in this case the measured execution time is greater than the estimated time and far from the the optimal one, thus minimizing the impact of the inaccuracy. The FMM-FFT implementation outperforms the FMM, except for the big cube size, where the FMM is slightly faster than the FMM-FFT. Once again, it is also worth mentioning that the execution time of the FMM with optimal cube size selection ($D/\lambda = 1.11$) is comparable to that of the FMM-FFT when using a small but suboptimal cube size ($D/\lambda = 0.5$).

Table 2: FMM and FMM-FFT parameters and execution times. ⌀2 m sphere analyzed at 8 kHz and discretized using 256020 unknowns.

$N = 256020$						
	FMM			FMM-FFT		
D/λ	N_g	$\widetilde{t}_{\text{it}}$	t_{it}	N_g	$\widetilde{t}_{\text{it}}$	t_{it}
0.5	36550	82.78 s	108.43 s	36550	14.65 s	16.60 s
0.82	-	-	-	14294	9.76 s	9.55 s
1.11	7916	19.84 s	19.25 s	-	-	-
2.5	1624	47.52 s	46.84 s	1624	46.73 s	46.95 s

Table 3: FMM and FMM-FFT parameters and execution times. ⌀2 m sphere analyzed at 16 kHz and discretized using 1023404 unknowns.

$N = 1023404$						
	FMM			FMM-FFT		
D/λ	N_g	$\widetilde{t}_{\text{it}}$	t_{it}	N_g	$\widetilde{t}_{\text{it}}$	t_{it}
0.5	146600	1304.48 s	1888.81 s	146600	117.40 s	146.10 s
0.91	-	-	-	47124	59.21 s	61.28 s
1.6	15534	146.68 s	149.39 s	-	-	-
2.5	6434	207.60 s	204.09 s	6434	189.01 s	188.45 s

Finally, Table 3 shows the parameters and execution times of a ⌀2 m sphere analyzed at 16 kHz and discretized using 1023404 unknowns. The cost of the optimal cube size selection is 2.06 s for the FMM and 3.45 s for the FMM-FFT. In this problem, just as the previous one, the estimated times are not much accurate when the cube size is small ($D/\lambda = 0.5$). However, the measured execution times are consistently greater than the estimated times and far from the the optimal ones, thus minimizing the impact of the inaccuracy. It is remarkable that the execution time of the FMM with the optimal cube size selection ($D/\lambda = 1.6$) is comparable to that of the FMM-FFT when using suboptimal cube sizes ($D/\lambda = 0.5$ and $D/\lambda = 2.5$).

 ISBN: 978-84-608-6082-2

4 Conclusions

In this paper, we have shown that the group size has a deep influence on the time cost that the single level FMM and the FMM-FFT achieve. In fact, the reduction of the time cost is up to one order of magnitude when the optimal D is considered. We have used octrees to perform a very efficient search of the optimal group size. This search is accomplished only once (during the setup) with a time cost that is negligible compared to the solver time. We anticipate that the technique employed to find the optimal group size may be applied to other versions of the algorithm (*Nested* FMM-FFT or even, MLFMM) and to other hardware architectures (such as GPUs).

Acknowledgements

This work has been supported by the "Ministerio de Economía y Competitividad" of Spain / FEDER under projects TEC2014-54005-P and TEC2015-67387-C4-3-R; and by the "Gobierno del Principado de Asturias" / FEDER under project FC-15-GRUPIN14-114.

References

[1] L. Greengard and V. Rokhlin, *A Fast Algorithm for Particle Simulations*, Journal of Computational Physics, **73**, (1987) 325–348.

[2] V. Rokhlin, *Diagonal Forms of Translation Operators for the Helmholtz Equation in Three Dimensions*, Applied and Computational Harmonic Analysis, **1**, (1993) 82–93.

[3] J. Song, C.-C. Lu and W. C. Chew, *Multilevel Fast Multipole Algorithm for Electromagnetic Scattering by Large Complex Objects*, IEEE Transactions on Antennas and Propagation, **45 (10)**, (1997) 1488–1492.

[4] C. Waltz, K. Sertel, M. A. Carr, B. C. Usner and J. L. Volakis, *Massively Parallel Fast Multipole Method Solutions of Large Electromagnetic Scattering Problems*, Transactions on Antennas and Propagation, **55 (6)**, (2007), 1810-1816.

[5] J. M. Taboada, L. Landesa, F. Obelleiro, J.L. Rodriguez, J. M. Bertolo, M. G. Araujo, J. C. Mourio, and A. Gomez, *High Scalability FMM-FFT Electromagnetic Solver for Supercomputer Systems*, IEEE Antennas and Propagation Magazine, **51 (6)**, (2009), 20-28.

[6] V. Dang, Q. M. Nguyen, and O. Kilic, *GPU Cluster Implementation of FMM-FFT for Large-Scale Electromagnetic Problems*, IEEE Antennas and Wireless Propagation Letters, **13**, (2014), 1259-1262.

 ISBN: 978-84-608-6082-2

[7] S. Scheneider, *Application of Fast Methods for Acoustic Scattering and Radiation Problems*, Journal of Computational Acoustics, **11 (3)**, (2003) 387–401.

[8] M. López-Portugués, J. A. López-Fernández, J. Ranilla, R. G. Ayestarán and F. Las-Heras, *Parallelization of the FMM on distributed-memory GPGPU*, Journal of Supercomputing **64 (1)** (2013) 17–27.

[9] K. Yoshida, N. Nishimura, S. Kobayashi, *Application of new fast multipole boundary integral equation method to crack problems in 3D*, Engineering Analysis with Boundary Elements **25** (2001) 239–247.

[10] R. Coifman, V. Rokhlin, and S. Wandzura, *The Fast Multipole Method for the Wave Equation: A Pedestrian Prescription*, IEEE Antennas and Propagation Magazine **35 (3)** (1993) 7–12.

[11] Nail A. Gumerov and Ramani Duraiswami and Eugene A.Borovikov, *Data Structures, Optimal Choice of Parameters, and Complexity Results for Generalized Multilevel Fast Multipole Methods in d Dimensions*, Institute for Advanced Computer Studies, 2003.

[12] J.A. López-Fernández, M.L. Portugués, J.M. Taboada, H.J. Rice and F. Obelleiro, *HP-FASS: a hybrid parallel fast acoustic scattering solver*, International Journal of Computer Mathematics, **88 (9)** (2011) 1960–1968.

[13] J. Song and W. Chew, *Multilevel Fast-Multipole Algorithm for Solving Combined Field Integral Equations of Electromagnetic Scattering*, Microwave and Optical Technology Letters, **10 (1)**, (1995), 14-19.

[14] R. Wagner, J. Song, and W. Chew, *Monte Carlo Simulation of Electromagnetic Scattering form Two-Dimensional Random Rough Surfaces*, IEEE Transactions on Antennas and Propagation, **45 (2)**, (1997), 1810–1816.

[15] M. G. Araújo, J. M. Taboada, F. Obelleiro, J. M. Bértolo, Luis Landesa, J. Rivero, J. L. Rodríguez, *Supercomputer aware approach for the solution of challenging electromagnetic problems*, Progress In Electromagnetics Research, **101**, (2010), 241-256.

[16] A. J. Burton and G. F. Miller, *The Application of Integral Equation Methods to the Numerical Solution of Some Exterior Boundary-Value Problems*, Proc. of the Royal Society of London, **323**, **1553**, (1971), 201–210.

[17] Y. Saad and M. H. Schultz, *GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*, SIAM J. of Sci. and Statist. Comput., **7**, (1986) 856–869.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A consistent first order theory about the equilibrium figures in close binary systems

José Antonio López Ortí[1], Manuel Forner Gumbau[1] and Miguel Barreda Rochera[1]

[1] *Departamento de Matemáticas, Universidad Jaume I de Castellón*

emails: lopez@mat.uji.es, fornerm@mat.uji.es, barreda@mat.uji.es

Abstract

In this paper the study of equilibrium figures of close binary systems is discussed. The classical theory on this subject is based on Laplace desideratum but, unfortunately, this hypothesis cannot be proved.

So, in this research a first order theory about this subject is offered without making use of Laplace desideratum. To do this two methods have been developed: the first one is based on the asymptotic properties of numerical quadrature and the second one is based on the Laplace method to determine the inverse of the distance between two planets.

The paper also contains algorithms to develop the product of the spherical harmonics by using Mathematica and a set of algorithms for managing the terms of the deformable body potential.

Key words: Celestial Mechanics. Figures of Celestial Bodies. Spherical Harmonics. Potential Theory. Close Binary Systems.

MSC 2000: 70F15, 70E20, 74G10, 76U05.

1 Introduction

The first aim of this work is to develop a consistent first order amplitudes theory to assess the self-gravitational potential of each component of a close binary system when the hydrostatic equilibrium has been achieved. This case can be modeled as:

$$\overrightarrow{\nabla} P = \rho \overrightarrow{\nabla} \Psi$$
$$\triangle \Psi = -4\pi G\rho + 2\omega^2$$

 ISBN: 978-84-608-6082-2

where P is the pressure, ρ is the density, Ψ is the total potential, $\triangle$ is *Laplace* operator, G is the gravitational constant, and $\overrightarrow{\omega}$ is the system's angular velocity.

To integrate these equations, in a general case of mass distribution, a state equation relating pressure with density is also needed.

To assess the full potential Ψ to calculate the self-gravitational potential Ω, the centrifugal potential V_c and the tidal potential V_t is needed. This hydrostatic equilibrium involves the rigid rotation of the system corresponding to the minimum potential. According with Kopal [4], this state involves identifying the equipotential, isobaric, isothermal and isopycnic surfaces.

To study the structure of the primary component, an then of the secondary, a coordinate system $OXYZ$ is defined, where O is the centre of mass of the primary component, OX is the axis in the direction of the center of mass of the secondary component, OZ an axis parallel to the angular velocity $\vec{\omega}$ of the system and OY is defined in a way that $OXYZ$ form a direct trihedron. So the Clairaut coordinates for an arbitrary point P of the primary component are expressed by (a, θ, λ) where a is the radius of the sphere with the same mass as the equipotential surface containing P and (θ, λ) is the angular part of the spherical coordinates of P.

The classical theory of this problem can be seen in Finlay [1], Kopal [4].

To achieve the aim of this research two different methods are proposed: the first one, which we will call analytical method, is similar to that used by Laplace to develop the inverse of the distance between two planets. The second one, named numerical quadrature method, is based on the asymptotic properties of the numerical quadrature formulas.

Furthermore, to obtain the series expansions of the inner and outer self-gravitational potential, we have deduced an algorithm to develop the product of spherical harmonics in real form as linear combination of themselves. For this we have relied on the theory and algorithms developed by Forner [2].

The main problem faced on this study is the development of the self-gravitational potential. To solve it we have followed two paths: The first one is based on the classic development of the potential as

$$\Omega = U + V, \qquad U = G\int_{r_0}^{r_1}\int_0^{2\pi}\int_0^{\pi}\frac{dm'}{\Delta}, \quad V = G\int_0^{r_0}\int_0^{2\pi}\int_0^{\pi}\frac{dm'}{\Delta}$$

where Δ is the distance between the element of mass dm' situated in P' and the point P, r_0 is the radius of the sphere centered on O containing P and r_1 is the radius of the smallest sphere centered at O containing the primary component. The element of mass is given by $dm' = \rho r'^2 \sin\theta' d\theta' d\lambda' dr'$. The second path is based on the evaluation of the self-gravitational potential given by

$$\Omega = G\int_V \frac{dm'}{\Delta} = G\int_{V_i}\frac{dm'}{\Delta} + G\int_{V_e}\frac{dm'}{\Delta}$$

where V_i is the inner part of the main component of the equipotential surface S_0 containing P and V_e is the region of the primary component contained between the equipotencial surface S_0 and the boundary of the S_e component.

2 Classical theory about the self-gravitational potential

The classical theory is based on the development of the inverse of the distance:

$$\frac{1}{\Delta} = \begin{cases} \frac{1}{r}\sum\limits_{n=0}^{\infty}\left(\frac{r'}{r}\right)^n P_n(\cos\gamma) & r > r' \\ \frac{1}{r'}\sum\limits_{n=0}^{\infty}\left(\frac{r}{r'}\right)^n P_n(\cos\gamma) & r < r' \end{cases}, \tag{1}$$

where γ is the angle between $\overrightarrow{OP}$ and $\overrightarrow{OP'}$. From this development we can write as $\Omega = \sum\limits_{n=0}^{\infty} U_n r^n + \sum\limits_{n=0}^{\infty} V_n r^{-n-1}$, where

$$U_n = G\int_{r_0}^{r_1}\int_0^{2\pi}\int_0^{\pi} \rho\, r'^{1-n} P_n(\cos\gamma)\sin\theta' dr' d\theta' d\lambda'.$$

$$V_n = G\int_0^{r_0}\int_0^{2\pi}\int_0^{\pi} \rho\, r'^{2+n} P_n(\cos\gamma)\sin\theta' dr' d\theta' d\lambda'.$$

The coordinate r is connected with the Clairaut coordinates (a,θ,λ) by

$$r = a\left(1 + \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f_{n,m}(a) Y_{n,m}(\theta,\lambda)\right)$$

where $f_{n,m}(a)$ are the amplitudes and $Y_{n,m}(\theta,\lambda)$ the spherical harmonics.

The classical theory [1], [4] assumes that

$$U_n = \frac{G}{2-n}\int_{a_0}^{a_1}\rho'\frac{\partial}{\partial a}\left[\int_0^{2\pi}\int_0^{\pi} r'^{2-n} P_n(\cos\gamma)\sin\theta' d\theta' d\lambda'\right] da', \quad if\ n \neq 2 \tag{2}$$

$$U_2 = G\int_{a_0}^{a_1}\rho'\frac{\partial}{\partial a}\left[\int_0^{2\pi}\int_0^{\pi} \ln(r') P_2(\cos\gamma)\sin\theta' d\theta' d\lambda'\right] da' \tag{3}$$

$$V_n = \frac{G}{n+3}\int_0^{a}\rho'\frac{\partial}{\partial a}\left[\int_0^{2\pi}\int_0^{\pi} r'^{n+3} P_n(\cos\gamma)\sin\theta' d\theta' d\lambda'\right] da' \tag{4}$$

Notice that these functions are defined in the internal and external region at the equipotential surface that contains P and it is necessary to assume that series converge in these regions [3], [5]. This assumption is known as Laplace desideratum.

From this hypothesis and approaching r'^k and $\ln r'$ up to the desiderate order in amplitudes we obtain $\Omega = 4\pi G\sum\limits_{n=0}^{\infty}\sum\limits_{m=-n}^{n}\frac{1}{2n+1}\left[E_{n,m}(a)r^n + F_{n,m}r^{-n-1}\right]Y_{n,m}(\theta,\lambda)$

 ISBN: 978-84-608-6082-2

3 A first order consistent theory of self-gravitational potential

In this section a technique to evaluate the self-gravitational potential is developed without using the Laplace desideratum. For this purpose we will use two different methods. The first one is based on the asymptotic properties of numerical quadratures and the second one on the Laplace development related to the inverse of the distance.

Let Σ be the equipotential surface containing P and $r = r_0$ the sphere containing P the equation of this sphere in Clairaut coordinates is given up to first order in amplitudes by $a' = a(1+D-D')$ where $D = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{n,m}(a) Y_{n,m}(\theta, \lambda)$ and $D' = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{n,m}(a') Y_{n,m}(\theta', \lambda')$. It is easy to notice that (2), (3) and (4) are false. However, up to the firsrt order the values obtained through the classical theory for de self-gravitational potential are correct.

A second independent way to prove this result is to evaluate the self-gravitational potential by means of the direct evaluation of potential Ω as:

$$\Omega = G \int_0^a \int_0^{\pi} \int_0^{2\pi} \frac{dm'}{\Delta} + G \int_a^{a_1} \int_0^{\pi} \int_0^{2\pi} \frac{dm'}{\Delta}$$

Let us defines $1/\Delta = \Psi(r, r') = 1/\sqrt{r^2 + r'^2 - 2rr' \cos\gamma}$. Developing $\Psi(r, r') = \Psi(a, a') + \Psi_{a'}(a, a') aD + \Psi_a(a, a') a' D'$ where Ψ_a, $\Psi_{a'}$ denotes partial derivative with respect a, a' and $\Psi(a, a')$ is given by (1). Operating we can obtain the same result.

4 Acknowledgments

This research has been partially supported by Grant AICO/2015/037 from the Generalitat Valenciana.

References

[1] E. Finlay-Frendulich, *Celestial Mechanics*, Pergamon Press Inc., New York 1958.

[2] M. Forner Gumbau, *Desarrollos en serie de los productos de algunas funciones especiales*, Publicacions de la Universitat Jaume I (Castellón - Spain), 2012.

[3] W. Jardetzky, *Theorie of figures of Celestial Bodies*, Interscience Publishers, Inc., New York, 1958.

[4] Z. Kopal, *Dynamic of Close Binary Systems*, Kluwer, Dordrecht, Holland 1978.

[5] F. F. Tisserand, *Traité de Mécanique Celeste*, Ed Gauthier-Villars, Paris, 1896.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Distributed Group Key Exchanges reusing randomness

J.A. López-Ramos[1], J. Rosenthal[2], D. Schipani[2] and R. Schnyder[2]

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

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

emails: jlopez@ual.es, rosenthal@math.uzh.ch, davide.schipani@math.uzh.ch, reto.schnyder@math.uzh.ch

Abstract

We introduce a key exchange scheme that allows a group of users to collaborate in order to share a common key. The scheme is formed by efficient protocols for both an initial key agreement and for rekeying.

Key words: Group Key Exchange, Diffie-Hellman Semigroup Action Problem
MSC 2000: AMS 95A60

1 Introduction

Security of group communications is an important topic of research given the numerous applications that provide information exchange in a community of users and/or devices, as is the case of the emerging Internet of Things, with numerous recent publications, e.g see [5], [10] and their references.

The first solution for Key Exchange was introduced by Diffie and Hellman in their foundational work [4] and since then many attempts have been proposed to extend this solution for a pair of users to a dynamic communication group where users are joining and leaving the group constantly, so that rekeying is a necessity to preserve confidentiality of communications. This is usually done in a distributed manner, i.e. users collaborate to build a shared key. Perhaps two of the best known schemes are due to Steiner et al. in [8] and [9], and to Burmester and Desmedt in [2] and [3]. These two proposals show how Diffie-Hellman key exchange can naturally be extended.

However, in [7] and [1], two active attacks are introduced for the two schemes, respectively, that allow the attacker to share a common key with the group without leaving any trace of his/her presence.

Our aim in this work is to introduce an efficient protocol for a distributed group key exchange that avoids these attacks while featuring both an efficient initial key agreement and a rekeying protocol. The scheme is presented in a general setting introduced in [6] considering group actions and it is based on the so-called Diffie-Hellman Semigroup Action Problem, DHSAP:

Problem ([6]) Given a finite abelian semigroup G acting on a finite set S and elements $x, y, z \in S$ with $y = g \cdot x$ and $z = h \cdot x$ for some $g, h \in G$, find $(gh) \cdot x$.

Throughout this paper G and H will denote abelian semigroups. We will say that an action $\Phi : G \times H \to H$ defined by $\Phi(g, s) = g \cdot s$ is linear in case $\Phi(g, ss') = g \cdot ss' = (g \cdot s)(g \cdot s') = \Phi(g, s)\Phi(g, s')$. Here we denote by yuxtaposition operations in both semigroups.

2 The Group Key Exchange

The general setting for the Group Key Exchange protocol consists in n participants that aim to share a common key built in a distributed manner. The set of participants or users will be given by $\{U_1, \dots U_n\}$. The set of users agree on a linear semigroup action $\Phi : G \times H \to H$ denoted by $\cdot$ and we will require the existence of inverses in H. They also agree on $h \in H$.

Every participant $\mathcal{U}_i$ holds two pairs of private-public keys, say $(g_i, g_i \cdot h)$ and $(x_i, x_i \cdot h)$. One of these users is chosen to be the group controller that we will denote by $\mathcal{U}_{c_1}$, for some c_1 in the set $\{1, \dots, n\}$. He will be in charge of sending the keying information. The protocol is given by the following steps.

Protocol 1

First Round:

1. Every user $\mathcal{U}_i$ publishes the pair $(g_i \cdot h, x_i \cdot h)$, $i = 1, \dots, n$, $i \neq c_1$.
2. The group controller $\mathcal{U}_{c_1}$ computes the key $K_1 = g_{c_1} \cdot \prod_{j=1, j\neq c_1}^{n} (g_j \cdot h)$.
3. The group controller takes two new elements $g'_{c_1}, x'_{c_1} \in G$ that become his new private information.

Second Round:

 ISBN: 978-84-608-6082-2

4. Every user $\mathcal{U}_i$, $i = 1, \ldots, n$, $i \neq c_1$, computes $\prod_{j=1, j\neq c_1, i}^{n} (g_j \cdot h)$ and sends this value to U_{c_1}.

5. The group controller $\mathcal{U}_{c_1}$ broadcasts the keying message

$$\{Y_{1,1}, \ldots, Y_{1,c_1}, \ldots, Y_{1,n}, R_1, S_1\}$$

where $Y_{1,i} = \left(g_{c_1} \cdot \prod_{j=1, j\neq c_1, i}^{n} (g_j \cdot h) \right) (x_{c_1} \cdot (x_i \cdot h))^{-1}$,
for $i = 1, \ldots, n$, $i \neq c_1$,
$Y_{1,c_1} = K_1 (g'_{c_1} \cdot (g_{c_1} \cdot h))^{-1} (x'_{c_1} \cdot (x_{c_1} \cdot h))^{-1}$,
and $R_1 = g_{c_1} \cdot h$ and $S_1 = x_{c_1} \cdot h$.

6. Every user $\mathcal{U}_i$ computes $K_{1,i} = Y_{1,i}(x_i \cdot S_1)(g_i \cdot R_1)$, $i = 1, \ldots, n$, $i \neq c_1$.

An immediate computation yields that at the end of the preceding protocol, all users hold the same key.

3 The Rekeying Protocol

The following protocol shows the Auxiliary Key Agreement after $t-1$ rekeying rounds, $t > 1$, whereby K_t denotes the last common key shared by the group. The user in charge of the t-th rekeying will be user $\mathcal{U}_{c_t}$, distinct from the preceding controller, and thus, rekeying information of this will be needed. Without loss of generality, we may assume that the precedent controller was user $\mathcal{U}_{c_1}$ and that the last rekeying message is given by

$$\{Y_{t-1,1}, \ldots, Y_{t-1,c_1}, \ldots, Y_{t-1,n}, R_{t-1}, S_{t-1}\}$$

being $Y_{t-1,c_1} = K_{t-1} (g'_{c_1} \cdot (g_{c_1} \cdot h))^{-1} (x'_{c_1} \cdot (x_{c_1} \cdot h))^{-1}$.

Protocol 2:

1. User $\mathcal{U}_{c_t}$ computes two new elements g'_{c_t} and $x'_{c_t} \in G$ that become his new private information.

2. User $\mathcal{U}_{c_t}$ computes the new session key $g'_{c_t} \cdot K_{t-1}$.

 ISBN: 978-84-608-6082-2

3. User $\mathcal{U}_{c_t}$ broadcasts the rekeying message

$$\{Y_{t,1}, \ldots, Y_{t,c_t}, \ldots, Y_{t,n}, R_t, S_t\}$$

where $Y_{t,i} = g'_{c_t} Y_{t-1,i}$, $i \neq c_t$,
$Y_{t,c_t} = K_t(g'_{c_t} \cdot (g'_{c_t} \cdot R_{t-1}))^{-1}(g'_{c_t} \cdot (x'_{c_t} \cdot S_{t-1}))^{-1}$,
and $R_t = g'_{c_t} \cdot R_{t-1}$ and $S_t = g'_{c_t} \cdot S_{t-1}$.

4. Every user $\mathcal{U}_i$ computes $K_{t,i} = Y_{t,i}(x_i \cdot S_t)(g_i \cdot R_t)$, $i = 1, \ldots, n$, $i \neq c_t$.

An induction argument shows that after running the rekeying protocol $t-1$ times, $t \geq 2$, every user holds the same key given by $K_{t+1} = g'_{c_t} K_{t-1}$, being K_1 the resulting key from *Protocol 1*.

Acknowledgements

The Research was supported in part by the Swiss National Science Foundation under grant No. 149716. First author is partially suppported by Ministerio de Economia y Competitividad grant MTM2014-54439 and Junta de Andalucia (FQM0211). The last author is supported by Armasuisse.

References

[1] M. Baouch, J.A. Lopez-Ramos, R. Schnyder and B. Torrecillas, *An active attack on a distributed Group Key Exchange system*, arXiv:1603.09090.

[2] M. Burmester and I. Desmedt, *A secure and efficient conference key distribution system*, in: Proc. Eurocypt'94, in: Lecture Notes in Comput. Sci., vol. 950, Springer-Verlag, Berlin, 1995, 275-286.

[3] M. Burmester and I. Desmedt, *A secure and scalable Group Key Exchange system*, Inf. Process. Lett. **94** (2005) 137–143.

[4] W.D. Diffie and M.E. Hellman, *New directions in cryptography*, IEEE Trans. Inform. Theory **22**(6) (1976) 644–654.

[5] P.P.C. Lee, J.C.S. Lui and D.K.Y. Yau, *Distributed Collaborative Key Agreement and Authentication Protocols for Dynamic Peer Groups*, IEEE ACM Trans. Network. **14**(2) (2006) 263–276.

[6] G. Maze, C. Monico and J. Rosenthal, *Public key cryptography based on semigroup actions*, Adv. Math. Commun. **1**(4) (2007) 489–507.

[7] R. Schnyder, J.A. Lopez-Ramos, J. Rosenthal and D. Schipani, *An active attack on a multiparty key exchange protocol*, J. Algebra Comb. Discrete Appl. **3**(1) (2016) 31–36.

[8] M. Steiner, G. Tsudik and M. Waidner, *Diffie-Hellman key distribution extended to group communication*, Proceedings of the 3rd ACM Conference on Computer and Communications Security, ACM: New York, NY, 31–37, 1996.

[9] M. Steiner, G. Tsudik and M. Waidner, *Key agreement in dynamic peer groups*, IEEE Tran. Parallel Distrib. Syst. **11**(8) (2000) 769–780.

[10] J. Van der Merwe, D. Dawoud and S. McDonald, *A survey on peer-to-peer key management for mobile ad hoc networks*, rm ACM Comput. Surv. 39 (1) 2007.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Parallel landing sites detection using LiDAR data on manycore systems

Oscar G. Lorenzo[1], Jorge Martínez[1], David L. Vilariño[1], Tomás F. Pena[1], José C. Cabaleiro[1] and Francisco F. Rivera[1]

[1] *CiTIUS Centro Singular de Investigación en Tecnoloxías da Información, Universidade de Santiago de Compostela*

emails: oscar.garcia@usc.es, jorge.martinez@usc.es, david.vilarino@usc.es, tf.pena@usc.es, jc.cabaleiro@usc.es, ff.rivera@usc.es

Abstract

Helicopters are widely used in emergency situations where knowing if a geographical location is adequate for landing is a critical issue, and it is far from being a straightforward task. In this work, we present an ongoing project to use LiDAR point clouds to detect and classify landing sites in real time. Landing sites are detected in parallel on manycore systems. Load balancing was identified as the main cause of poor performance. We present results for a set of LiDAR point clouds and balancing strategies for three different multi and manycore systems. The load balancing techniques applied increase performance up to 3 times from the unbalanced case.

Key words: LiDAR, landing, load balancing, Xeon Phi

1 Introduction

Helicopters are widely used in emergency situations. They are responsible for rescue operations or firefighting efforts, sometimes in difficult terrain areas. Under these situations, knowing if a place is adequate for landing is not a straightforward task, even for experienced pilots. In these cases, having in advance a map with possible landing sites, or being able to create one on real time, is a clear advantage. This map may be created using a camera or, in 3D, by a LiDAR system, any of which can be installed in a helicopter. In this work, we present an ongoing project to use LiDAR point clouds to detect and classify landing sites in real time..

LiDAR has become an established method for collecting accurate information of the earth landscape. The record of elevation data and the ability to penetrate through the canopy, are some of the advantages over traditional adquisition methods like aerial imagery.

The equipment consists of an airborne laser scanner (ALS), an inertial measurement unit (IMU) and a GPS receiver. These devices are installed in an aircraft that flies over the ground surface obtaining a georeferenced point cloud. On land it is also a GPS system synchronized with the GPS receiver on the aircraft. For each point a set of features are recorded: x and y coordinates, elevation, intensity, pulse id, pulses number and scan angle.

LiDAR data has been proven useful for landing hazard detection for helicopters [1] or even spaceships [2], using, for example, the incidence angle of the rays and the roughness of the terrain. It has been used for autonomous landing in real time, using geometrical approaches [3] or neural networks [4] to detect the quality of the site.

Nevertheless, parallelising LiDAR data processing may be complex [5], but it is needed to achieve real time performance. Building recognition [6], road detection [7] or object recognition [8] have been shown to be parallel tasks. Still, one of the issues it faces is load balancing, specially in hybrid systems [9].

In this paper, we show how the workload may be balanced in a set of systems, including manycores were parallelisation is essential [10] to achieve efficient parallel solutions to the landing problem hat are needed in real time situations. The rest of the paper is organised as follows: in Section 2 the landing detection algorithm is described. In Section 3, a Case Study is presented, with examples of the landing detection, a comparison execution times in 3.1 and the effect of load balancing in 3.2. Finally, conclusions are in Section 4.

2 Landing Problem

Pilots need a classification of all the points in the terrain where an helicopter could be landed. This should be done as quick as possible, so it can be used in real time emergency operations. A landing site should fulfill some of the following requirements, as indicated by a professional pilot from INAER (www.inaer.com):

- A planar surface where the helicopter skids may land:
  - Must be large enough for the skids.
  - Must have a slope not steeper than a given threshold.
  - Vegetation and other obstacles in the surface must not exceed a given height.
- An area devoid of obstacles large enough for the helicopter rotors and fuselage.

For each of these requirements different thresholds may be considered in order to classify the suitability for landing; for example, to detect landing zones of a greater landing difficulty, acceptable for an emergency, or good landing zones.

 ISBN: 978-84-608-6082-2

The point cloud is stored in an octree for its efficient processing. An octree is a tree structure in which each node has exactly eight children. The space is partitioned recursively until the radius of a octant is below a given threshold, 2 m in this work. The use of this octree structure speeds up considerably the search of neighbors of a point [11].

The landing detection algorithm works on each LiDAR point individually, as follows:

First, a planar surface is calculated centered in the point. The plane of a set of points is calculated using Principal Component Analysis (PCA). That is, the plane of a point is estimated by performing PCA in his neighborhood. The neighborhood of each point is defined as a square centered in the point, in this case with a side length dependent on the selected skid size. All the LiDAR points in this neighborhood are searched in the octree. The covariance matrix is then calculated and the eigenvectors are extracted. The normal vector of the point is the eigenvector that yields the smallest eigenvalue. This plane, and the distances of the neighbor points to it, are enough to classify a point as unsuitable, and the same plane is then used to calculate the distance to the ground of all the points in the possible landing site. So, according to the requirements for skid landing –such as slope or vegetation– the point is classified as unsuitable, when any requirement is not fulfilled, or the algorithm continues to check other variables of the landing site. Thus, if the point is not unsuitable, the algorithm continues searching if the larger area around it can accommodate the rotors and fuselage.

Next, a new neighborhood is searched, this time with radius equal to the rotor's. According to the points in this larger area, the point can be classified as planar –when the skids may land, but not enough space for the rotors is accounted for–, or a larger area is searched to further decide the quality of the landing site. Since this neighborhood is far larger, comprising many more points, it takes a longer time to gather. These distances are enough to classify a point as planar or a possible landing site.

Finally, the quality of the landing site is determined. To check if the landing site is acceptable or good, a new neighborhood, this time with a radius equal to the rotor's and fuselage, plus a safety distance, must be searched. Checking the distances of all these neighborhood points to the plane determines if the landing site is finally risky –when landing is possible but risky–, acceptable or good. In this way, each point is fully classified in a single iteration. Therefore, there is no need for computing the whole point cloud to find a good enough solution. A flow diagram of this algorithm is show in Figure 1.

The points in the neighborhoods being studied are classified temporally as belonging to a planar, risky, acceptable or good landing site, if that is the case. This temporal classification may be overridden if they become the center of a landing site.

The process of gathering all these neighborhoods means that not all the points are classified at the same time, since an unsuitable point is already classified without needing to check the larger areas. Note that, a planar point does not require calculating the last neighborhood. Moreover, the number of points inside these neighborhoods may vary greatly.

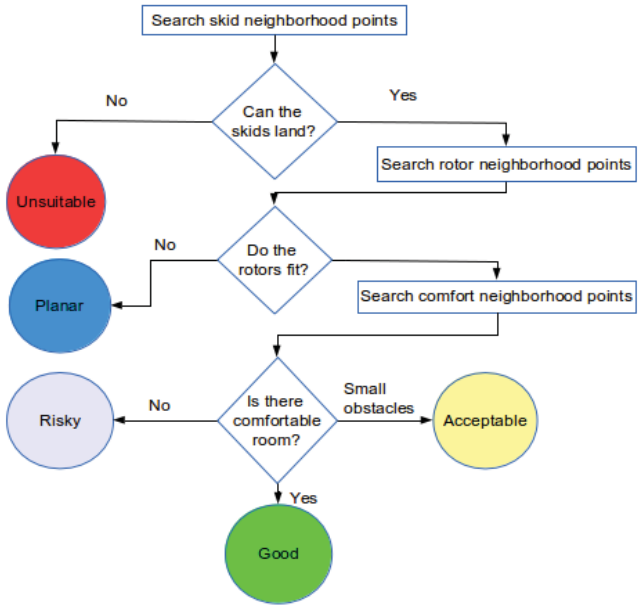


Figure 1: Landing algorithm flow diagram.

This fact generates a load balancing issue when the algorithm is parallelised, even while all points are processed independently. Nevertheless, this algorithm, as has been previously stated, allows for any landing point to be classified in one sweep, which is useful when working with changing data.

The load balance was identified as the main factor to improve performance. To solve this issue the algorithm was parallelised with OpenMP [12] using load balancing clauses. First, to parallelise the algorithm, each thread keeps a copy of all the points classification as an array, and in the end this classification array is reduced to the best classification for each point. In this way all the threads can work independently and the main loop can be parallelised with OpenMP, and balanced with the `schedule(dynamic,block_size)` clause, where `block_size` is the number of points given to each thread at a time. With this configuration, the OpenMP runtime splits the number of points in blocks of size `block_size`, and assigns each block to one thread. As soon as a thread finishes its work on its allotted block, a new block is assigned to it, until all the blocks are processed.

Table 1: Percentage of classified points.

	Points classified as (in %)			
map	unsuitable	planar	risky	acceptable or good
Guitiriz	64.75	33.10	>0.01	2.14
Alcoy	53.64	43.81	0.15	2.41
Vaihingen	40.94	50.55	0.68	7.83
Australia	30.83	39.40	2.35	27.43
Florida	30.27	41.62	2.27	25.83
Australia Large	43.58	40.19	1.56	14.67
Florida Large	24.47	59.01	1.82	14.70

3 Case Study

To study the classification algorithm, 5 LiDAR maps with 1,700,000 points –so called Guitiriz, Alcoy, Vaihingen, Australia and Florida–, and 2 more with 5,000,000 points –Australia Large and Florida Large–, were selected. Guitiriz and Alcoy are private datasets, while the rest are public: Vaihingen dataset was downloaded from the ISPRS Benchmark Dataset [13], and Australia and Florida datasets were downloaded from OpenTopography [14]. Each of these maps results in a different ratio of classified landing points, as shown in Table 1, resulting in different execution times and load balancing issues.

In the particular case study here described, based on a series of thresholds indicated by INAER pilots, points are classified, and shown in the figures, as follows:

- Unsuitable, shown in red: The site is not adequate for landing.
- Planar, shown in blue (with points in the corresponding neighborhood in dark blue, this points are called accompanying points): The site is only large enough for the skids, that is a planar 4x4 m^2 square with less than a 10% slope and vegetation or obstacles lower than 50 cm.
- Risky, shown in white (with accompanying points in light blue): The site is risky for landing, rotor space is narrow, only 22 m.
- Acceptable, shown in yellow (with accompanying points in light yellow): The site is acceptable for landing, rotor space is at least 25 m, but the slope is between 8% and 10% and there are some obstacles between 30 cm and 50 cm.

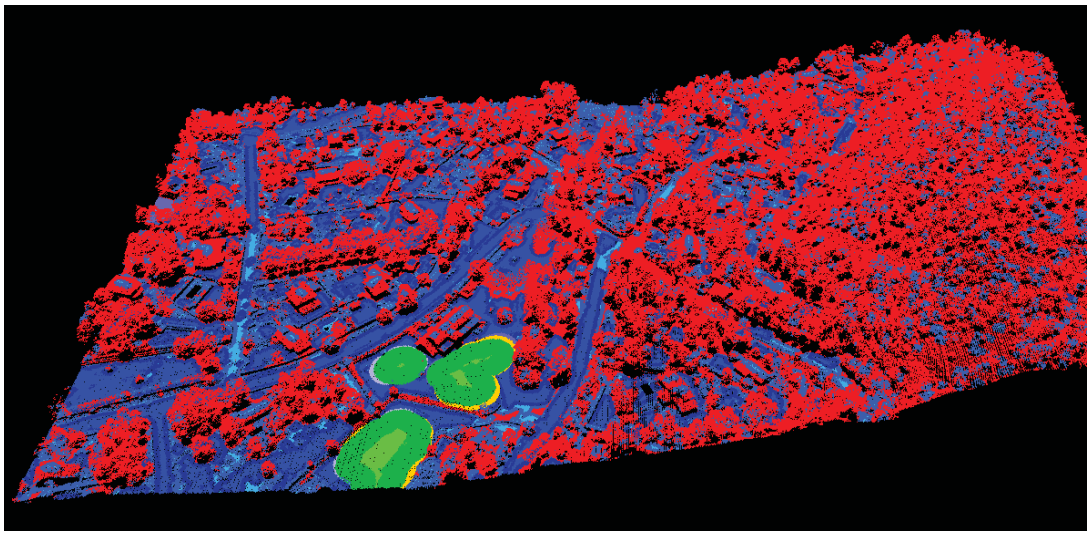


Figure 2: Detected landing points in Alcoy.

- Good, shown in light green (with accompanying points in dark green): The site is good for landing, rotor space is at least 25 m, the slope is below 8% and any obstacles or vegetation are below 30 cm.

An example of the classification for the Alcoy map is shown in Figure 2. In this example, many points are in a tree filled hill –on the left side of the figure– and are classified as unsuitable (in red). A few roads show as planar points (blue) and three landing sites (green), of different sizes, are found in a field and around a house.

Another example, that is part of the Australia Large map, is shown in Figure 3. In this sparsely populated area, many landing areas, some only acceptable (yellow), are found in various fields, with a few risky areas (light blue) between buildings –f.i. on the left hand side–.

3.1 Execution Times

The classification algorithm was executed in three different systems:

- A regular desktop -henceforth called Core-, with a Linux 3.2.0-97-generic Ubuntu, an Intel Core2 Quad CPU Q9550 @2.83GHz with 4 cores and 8 GB RAM.

- A manycore processor -henceforth called Phi-, with a Linux 2.6.38.8+mpss3.5 and a Xeon Phi 7120P with 61 cores and 244 threads.

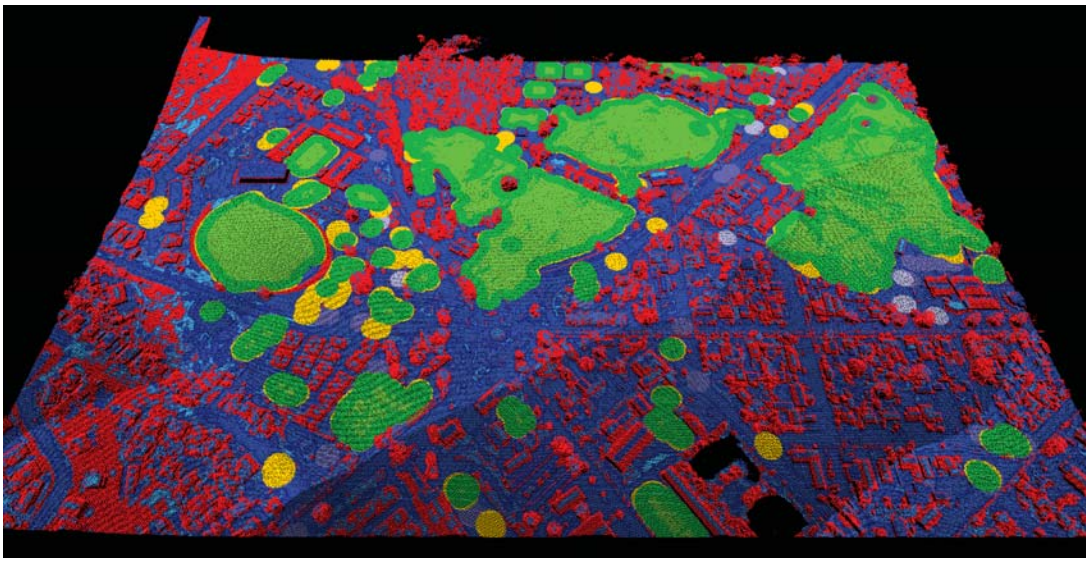


Figure 3: Some of the detected landing points in Australia.

- A powerful desktop -henceforth called Xeon-, with a Linux 3.13.0-79-generic Ubuntu, an Intel Xeon CPU E5-1650 @3.20GHz with 6 cores, 12 threads and 8 GB RAM.

The code was compiled with gcc-4.6.3 and O2 on the Core and Xeon and icc-15.0.2 and O2 on the Phi. The data input and output execution time is directly proportional to the number of points and does not vary much with the map –only with the system–, so only the execution time of the main landing points detection loop is shown from here onward. To test the load balance, the parameter `block_size` may be set for each execution. In Figure 4, the execution times for the best value of `block_size` for each map is shown for all the systems. This figure shows that the best results are obtained on the Xeon -except on the Vaihingen map-, with the Core performing much poorly, as expected. It also shows the input map clearly influences the execution time.

3.2 Load balance

As stated, load balance was tested varying the `block_size` parameter. With this parameter the total number of blocks can be determined if the number of threads is also known. To study the load balance the number of blocks per thread is varied from 2, where all threads should execute one or two blocks, to $number\_of\_points/threads$ where each thread should classify one point at a time. The case when all threads execute just 1 block is performed

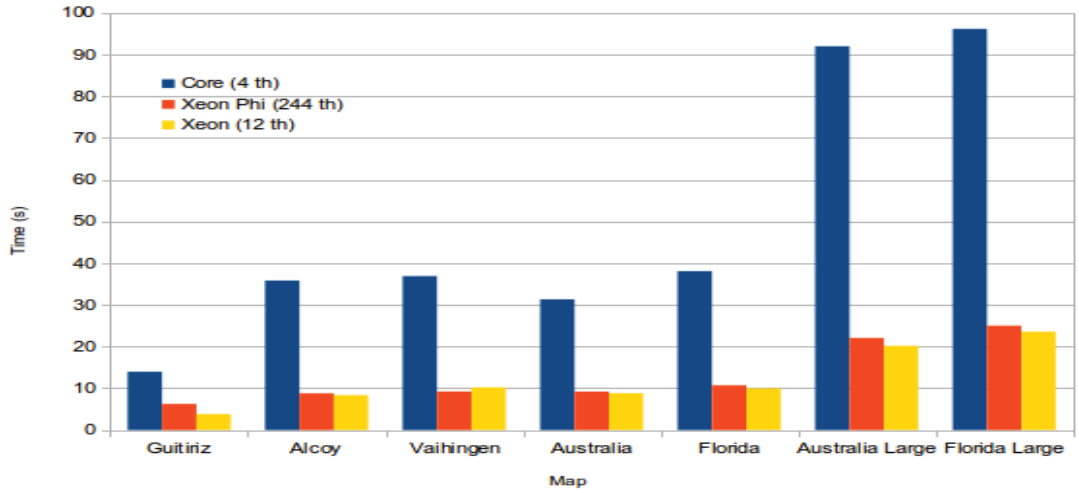
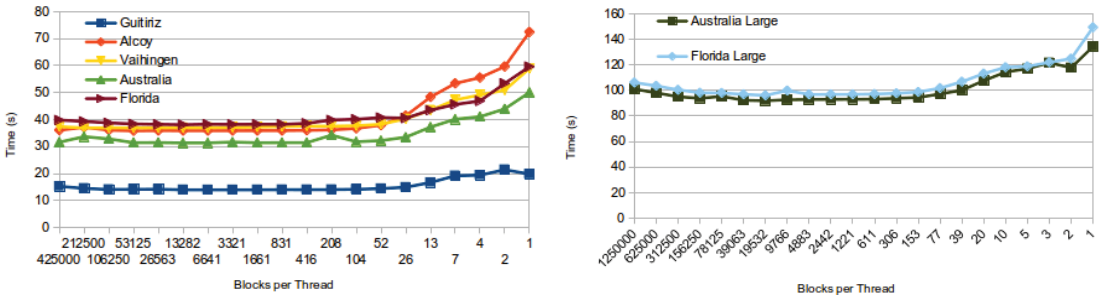


Figure 4: Execution times for the best `block_size` on all processors.



(a) 1,700,000 points.

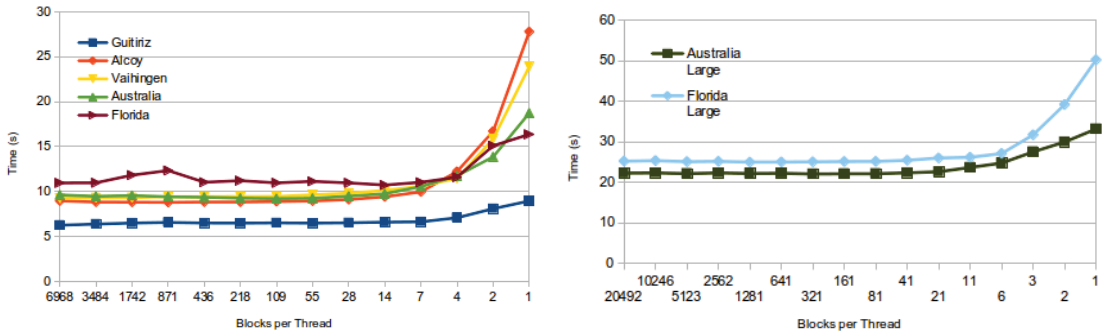
(b) 5,000,000 points.

Figure 5: Execution times on the Intel Core with 4 threads.

using OpenMP without scheduling clauses, so no scheduling is done and all work is divided beforehand.

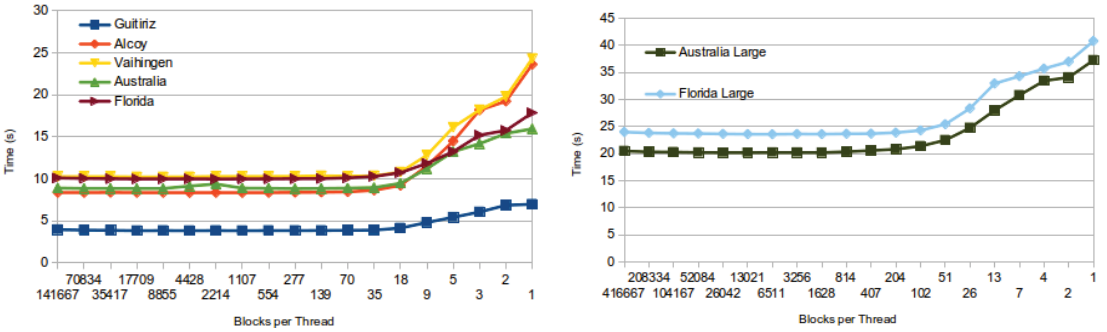
The results for the Core system with 4 threads are shown in Figures 5(a) and 5(b). These results show that, once the work is split in more than 20 blocks per thread, the workload balances, although the best results are with a few thousand blocks.When the work is not balanced, execution times are often about 2 times worse than the well balanced cases.

In Figures 6(a) and 6(b) the results for the Phi system are shown. In this system the work balances from about 5 blocks per processor. When the workload is unbalanced, results



(a) 1,700,000 points. (b) 5,000,000 points.

Figure 6: Execution times for large maps on the Xeon Phi with 244 threads.



(a) 1,700,000 points. (b) 5,000,000 points.

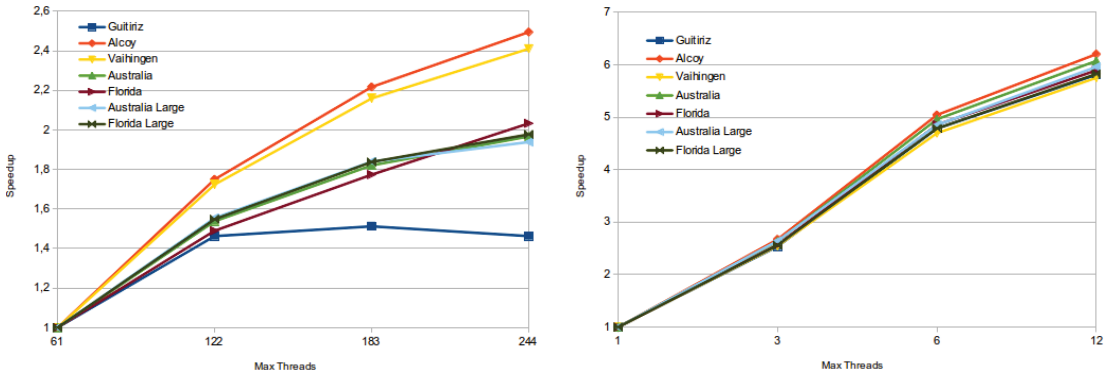
Figure 7: Execution times for large maps on the Xeon with 12 threads.

are 2 to 3 times worse, due to the importance of parallelism in the manycore architecture.

Results for the Xeon system can be seen in Figures 7(a) and 7(b). In this case the load balances from around 15 blocks per thread, and the gain against the unbalance case is about 2.5 times.

Taking into account the best load balanced cases, the speedup for the Phi system –with baseline in one thread per core and all cores used– is shown in Figure 8(a). While the speedup does not reach 4, it remains good as more threads are added per core. An exception is the Guitiriz map, which is already quite fast and does not benefit from extra threads.

In Figure 8(b) the speedup for the Xeon system is shown –with baseline in one thread and one core–. Again, the speedup does not reach the optimal of 12, but remains good,



(a) Xeon Phi. (b) Xeon.

Figure 8: Speedups using the best execution times.

taking advantage of the multithreading when there are 2 threads per core.

4 Conclusions

In emergency situations, knowing if a geographical position is adequate for an helicopter to land is not straightforward. We are developing a landing site detection tool that, working with LiDAR data, can classify landing points in real time. We are testing its usefulness and precision in a series of maps with different characteristics. We achieve real time situations execution times thanks to parallelising the application, which can run in commodity hardware. Load balance was detected as the main obstacle to achieve good performance and it was characterised. Three different systems were considered with 4, 6 and 61 cores. The load balancing techniques applied increase performance up to 3 times from the unbalanced case.

Acknowledgements

This work has been partially supported by the Ministry of Ministry of Economy and Competitiveness of Spain under project TIN2013-41129-P and Xunta de Galicia under projects GRC2014/008 and GRC GI-1638. It has been developed in the framework of the European network HiPEAC-2, the Spanish network CAPAP-H, the Galician network under the Consolidation Program of Competitive Research Units (TLIX Network ref. R2014/049). This work is also a result of a collaboration with INAER.

References

[1] M. Whalley, M. Takahashi, P. Tsenkov, G. Schulein, and C. Goerzen, "Field-testing of a helicopter UAV obstacle field navigation and landing system," in *65th Annual Forum of the American Helicopter Society, Grapevine, TX*, 2009.

[2] A. E. Johnson, A. R. Klumpp, J. B. Collier, and A. A. Wolf, "LiDAR-based hazard avoidance for safe landing on Mars," *Journal of guidance, control, and dynamics*, vol. 25, no. 6, pp. 1091–1099, 2002.

[3] S. Scherer, L. Chamberlain, and S. Singh, "Autonomous landing at unprepared sites by a full-scale helicopter," *Robotics and Autonomous Systems*, vol. 60, no. 12, pp. 1545–1562, 2012.

[4] D. Maturana and S. Scherer, "3D convolutional neural networks for landing zone detection from LiDAR," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3471–3478.

[5] J.-H. Jung, M. M. Crawford, and S.-H. Lee, "Complexity estimation based work load balancing for a parallel LiDAR waveform decomposition algorithm," *Korean Journal of Remote Sensing*, vol. 25, no. 6, pp. 547–557, 2009.

[6] H. J. Lee, "Parallel algorithm for building extraction from LiDAR data," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012, p. 1.

[7] J. Li, H. J. Lee, and G. S. Cho, "Parallel algorithm for road points extraction from massive LiDAR data," in *Parallel and Distributed Processing with Applications, 2008. ISPA'08. International Symposium on*. IEEE, 2008, pp. 308–315.

[8] J. Gong, H. J. Lee, and G. Cho. Parallel algorithm for object extraction from ranging data. [Online]. Available: http://dcs.chonbuk.ac.kr/~jgong/papers/Parallel%20Algorithm%20for%20Object%20Extraction%20from%20Ranging%20Data.pdf

[9] T. Zhang and J. Li, "Online task scheduling for LiDAR data preprocessing on hybrid GPU/CPU devices: A reinforcement learning approach," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 8, no. 1, pp. 386–397, 2015.

[10] A. Ramachandran, J. Vienne, R. Van Der Wijngaart, L. Koesterke, and I. Sharapov, "Performance evaluation of NAS parallel benchmarks on intel Xeon Phi," in *Parallel Processing (ICPP), 2013 42nd International Conference on*. IEEE, 2013, pp. 736–743.

[11] A. S. M. Mosa, B. Schön, M. Bertolotto, and D. F. Laefer, "Evaluating the benefits of octree-based indexing for LiDAR data," *Photogrammetric Engineering & Remote Sensing*, vol. 78, no. 9, pp. 927–934, 2012.

[12] L. Dagum and R. Enon, "OpenMP: an industry standard api for shared-memory programming," *Computational Science & Engineering, IEEE*, vol. 5, no. 1, pp. 46–55, 1998.

[13] ISPRS. (2016) Isprs test project on urban classification, 3d building reconstruction and semantic labeling. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/tests.html

[14] OpenTopography. (2016) Lidar point cloud data distribution and processing. [Online]. Available: http://opentopo.sdsc.edu/lidar

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Stop&Restart vs Resilient MPI applications

Nuria Losada[1], María J. Martín[1] and Patricia González[1]

[1] *Computer Architecture Group, Universidade da Coruña, Spain*

emails: { nuria.losada, mariam, patricia.gonzalez }@udc.es

Abstract

Future exascale systems are predicted to be formed by millions of cores. This is a great opportunity for HPC applications, however, it is also a hazard for the completion of their execution. These machines will present higher failure rates, and long-running applications will need to use fault tolerance techniques to ensure the finalization of their execution. Despite being one of the most popular parallel programming models, the MPI standard lacks fault tolerance support and, traditionally, failures are addressed with stop&restart checkpointing solutions. The proposal of ULFM (User Level Failure Mitigation) for the inclusion of resilience capabilities in the MPI standard provides new opportunities in this field, allowing the implementation of resilient MPI applications, i.e. applications that are able to detect and react to failures without stopping their execution. This work compares the performance of a traditional stop&restart checkpointing solution with its equivalent resilience proposal. The evaluation is focussed on the scalability of both solutions, assessing the proposals using up to 3027 cores. Preliminary results show that the resilience proposal outperforms the traditional stop&restart solution.

Key words: Resilience, Checkpointing, Fault Tolerance, MPI

1 Introduction

Current petascale systems are formed by hundreds of thousands of cores. Di Martino et al. [5] have studied during 518 days the Cray supercomputer Blue Waters, reporting that 1.53% of applications running on the machine failed because of system-related issues. This means that, on average, a failure arises every 15 minutes. Overall, failed applications noticeably run for about 9% of the total production node hours. The electricity cost of not using any fault tolerance mechanism in the failed applications was estimated at almost half

a million dollars during the studied period of time. Future exascale systems will be formed by several millions of cores, and they will be hit by error/faults much more frequently than petascale systems due to their scale and complexity [3]. Therefore, long-running applications in these systems will need to use fault tolerance techniques to ensure the completion of their execution and to save energy.

The MPI (Message Passing Interface) standard is the most popular parallel programming model in petascale systems. However, MPI lacks fault tolerance support. By default, the entire MPI application is aborted upon a single process failure. Besides, the state of MPI will be undefined in the event of a failure and there are no guarantees that the MPI program can successfully continue its execution. Thus, traditional fault tolerant solutions for MPI applications rely on stop&restart checkpointing: the computation state is periodically saved to stable storage into checkpoint files, so that, when a failure occurs, the application can be relaunched and its state recovered. However, when a failure arises it frequently has a limited impact and affects only a subset of the cores or computation nodes in which the application is being run. Thus, most of the nodes will still be alive. In this context, aborting the MPI application to relaunch it again introduces unnecessary recovery overheads and more efficient solutions need to be explored.

In the last years new methods have emerged to provide fault tolerance to MPI applications, such as failure avoidance approaches [4, 9] that preemptively migrate processes from processors that are about to fail. Unfortunately, these solutions are not able to cope with already happened failures. Recently, the Fault Tolerance Working Group within the MPI forum proposed the ULFM (User Level Failure Mitigation) interface [2] to integrate resilience capabilities in the future MPI 4.0 standard. It includes new semantics for process failure detection, and communicator revocation and reconfiguration. Thus, it enables the implementation of resilient MPI applications, that is, applications that are able to recover themselves from failures.

CPPC [8] is an open-source checkpointing tool for MPI applications, which originally applies a stop&restart checkpointing strategy. In a previous work [7] CPPC has been extended to transparently obtain resilient MPI applications from generic MPI SPMD (Single Program Multiple Data) programs by exploiting the new resilience capabilities provided by ULFM. The CPPC resilience proposal is able to detect failures in one or multiple processes, and to recover from them, without stopping the execution of the MPI application. In this work, the CPPC resilience proposal is exhaustively evaluated in a large machine (using up to 3072 cores) considering different fault scenarios to analyze the scalability of the proposal. Besides, it is compared with the traditional stop&restart checkpointing strategy in terms of performance, benefits and possible bottlenecks.

2 Experimental results

The experimental evaluation is performed at CESGA (Galicia Supercomputing Center, Spain) in the new FinisTerrae-II supercomputer, comprised of nodes with two Intel Xeon E5-2680 v3 @ 2.50GHz processors, with 12 cores per processor and 128 GB of RAM, interconnected to an InfiniBand FDR 56Gb/s. The performance of both the stop&restart and the resilience proposal is evaluated in two different fault scenarios: inserting one-process or full-node failures. One-process failures are simulated by killing one of the MPI processes executing the application, while in full-node failures all the MPI processes running in a computation node are killed.

The application testbed used is comprised of three benchmarks with different checkpoint file sizes and communication patterns: the ASC Sequoia Benchmark SPhot [1] (a physics package that implements a Monte Carlo Scalar Photon transport code), the Himeno benchmark [6] (a Poisson equation solver using the Jacobi iteration method) and the MOCFE-Bone [10] application (which simulates the main procedures in a 3D method of characteristics (MOC) code for numerical solution of the steady state neutron transport equation).

Preliminary results account for a superior performance of the resilience proposal. Although the overhead in the absence of failures in the resilience proposal is slightly superior to the one obtain with the stop&restart solution, the entire resilience recovery process is, on average, 3.2 times faster than in the stop&restart proposal. The benefits of the resilience proposal include a faster failure detection and reading of the checkpoint files. Also, the time to spawn only the failed processes is inferior to the time spent when relaunching the entire MPI application.

Acknowledgements

This work has been supported by the Ministry of Economy and Competitiveness of Spain and FEDER funds of the EU (project TIN2013-42148-P, predoctoral grant of Nuria Losada ref. BES-2014-068066, and CAPAP-H5 network TIN2014-53522-REDT) and by the Galician Government (Xunta de Galicia) under the Consolidation Program of Competitive Research Units, cofunded by FEDER funds of the EU (Ref. GRC2013/055). We gratefully thank CESGA (Galicia Supercomputing Center, Santiago de Compostela, Spain) for providing access to the FinisTerrae-II supercomputer.

References

[1] ASC Sequoia Benchmark Codes. `https://asc.llnl.gov/sequoia/benchmarks/`. Last accessed: September 2015.

[2] W. Bland, A. Bouteiller, T. Herault, J. Hursey, G. Bosilca, and J.J. Dongarra. An evaluation of user-level failure mitigation support in MPI. In *Recent Advances in the Message Passing Interface*, volume 7490 of *Lecture Notes in Computer Science*, pages 193–203. Springer, 2012.

[3] F. Cappello. Fault tolerance in petascale/exascale systems: Current knowledge, challenges and research opportunities. *International Journal of High Performance Computing Applications*, 23(3):212–226, 2009.

[4] I. Cores, G. Rodríguez, P. González, and M.J. Martín. Failure avoidance in MPI applications using an application-level approach. *The Computer Journal*, 57(1):100–114, 2014.

[5] C. Di Martino, W. Kramer, Z. Kalbarczyk, and R. Iyer. Measuring and Understanding Extreme-Scale Application Resilience: A Field Study of 5,000,000 HPC Application Runs. In *IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 25–36, 2015.

[6] Himeno Benchmark. `http://accc.riken.jp/2444.htm`. Last accessed: September 2015.

[7] N. Losada, I. Cores, M. J. Martín, and P. González. Resilient mpi applications using an application-level checkpointing framework and ulfm. *The Journal of Supercomputing*, pages 1–14, 2016. ISSN 1573-0484. doi: 10.1007/s11227-016-1629-7. URL `http://dx.doi.org/10.1007/s11227-016-1629-7`.

[8] G. Rodríguez, M.J. Martín, P. González, J. Touriño, and R. Doallo. CPPC: a compiler-assisted tool for portable checkpointing of message-passing applications. *Concurrency and Computation: Practice and Experience*, 22(6):749–766, 2010.

[9] C. Wang, F. Mueller, C. Engelmann, and S.L. Scott. Proactive process-level live migration in HPC environments. In *ACM/IEEE conference on Supercomputing*, pages 1–12, 2008.

[10] E. Wolters and M. Smith. MOCFE-Bone: the 3D MOC mini-application for exascale research. Technical report, Argonne National Laboratory (ANL), 2013.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A deterministic model for the distribution of the stopping time in a stochastic model and its numerical solution

Jorge Eduardo Macías-Díaz[1] **and José Villa-Morales**[1]

[1] *Departamento de Matemáticas y Física, Universidad Autónoma de Aguascalientes*

emails: jemacias@correo.uaa.mx, jvilla@correo.uaa.mx

Abstract

Departing from a general stochastic differential equation with Brownian diffusion, we establish that the distribution of probability of the stopping time is governed by a nonlinear parabolic partial differential equation with Dirichlet boundary conditions. An implicit, convergent and probability-based discretization to approximate the solution of the boundary-value problem is proposed in this work. Using a convenient vector representation of our scheme, we prove that the method preserves the most relevant properties of a probability distribution function, namely, the non-negativity, the boundedness from above by 1, and the monotonicity. *Key words: stochastic differential equations, nonlinear partial differential equations, probability distribution of hitting time, probability-based numerical method*

1 Introduction

Throughout, we suppose that b and σ are real functions which are continuously differentiable in a suitable unbounded interval of $\mathbb{R}$, and that W is a Brownian motion defined on a probability space $(\Omega, \mathcal{F}, P)$. Assume also that $x \in \mathbb{R}$, and that $t \geq 0$ physically represents time. The model of interest in this work is described by the stochastic differential equation

$$X_t = x + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s, \quad t > 0. \tag{1.1}$$

It is a well-known fact [1, Section 3.3] that for each $x \in \mathbb{R}$ there exists a unique stochastic process X defined up to a stopping time $e_x = \inf\{t : |X_t| = +\infty\}$, such that X is continuous on $[0, e_x)$ and satisfies (1.1) almost surely on $[0, e_x)$. This process is denoted by X^x, and the extended real number e_x is called the *explosion time* or the *blow-up time* of X^x. We say that X^x *explodes in finite time* when the explosion time is real.

 ISBN: 978-84-608-6082-2

Proposition 1.1. *Suppose that b and σ are continuously differentiable in some interval $[x_0, \infty) \subset \mathbb{R}$, and satisfy $b(x_0) \geq 0$ and $\sigma(x_0) = 0$. If $x \in (x_0, \infty)$ then $X_t^x > x_0$ for all $t > 0$ almost surely.* □

As a consequence, if X^x explodes in finite time then it must blow up towards $+\infty$. In the following, we study the distribution of the *stopping time at $\xi > x_0$*, which is defined as

$$\tau_\xi^x = \inf\{t \geq 0 : X_t^x \geq \xi\}. \tag{1.2}$$

Under the hypotheses of Proposition 1.1, the probability distribution of the stopping time at ξ satisfies $P[\tau_\xi^x \leq t] = 1$ for each $x \geq \xi$ and each $t > 0$, so it only remains to determine the probability distribution for $x_0 < x < \xi$. The following is a useful tool in the solution of that problem, and also the most important analytical result of this work.

Theorem 1.2. *Let b and σ be continuously differentiable in some $[x_0, \infty) \subset \mathbb{R}$, and such that $b(x_0) \geq 0$ and $\sigma(x_0) = 0$. Let $\xi > x_0$, and assume that $u : [0, \infty) \times (x_0, \infty) \to \mathbb{R}$ is a bounded function that satisfies the boundary-value problem*

$$\begin{aligned} &\frac{\partial u}{\partial t}(t,x) = Au(t,x), \quad (t,x) \in (0,\infty) \times (x_0, \infty), \\ &\quad \textit{subject to } \begin{cases} u(0,x) = 0, & x \in (x_0, \xi), \\ u(t,\xi) = 1, & t \in [0, \infty), \end{cases} \end{aligned} \tag{1.3}$$

where the differential operator A is given by

$$A = \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2} + b(x)\frac{\partial}{\partial x}. \tag{1.4}$$

Then $u(t,x) = P\left[\tau_\xi^x \leq t\right]$ for each $x_0 < x < \xi$ and each $t > 0$. □

2 Preliminaries

The phenomenon of crack propagation in materials due to fatigue is an important topic of investigation in structural mechanics [2]. In general, the process of propagation of a crack in a solid spans through three regions which depend on the length of the crack, and for which the dynamics of propagation is substantially different:

I. **Threshold.** The crack is imperceptible and its length is microscopic. In this stage, the microscopic structure of the material plays an important role.

II. **Paris' region.** The crack is usually visible and increases notably.

III. **Fracture.** The length of the crack increases abruptly and the fracture occurs all of a sudden.

 ISBN: 978-84-608-6082-2

The modeling of crack growth in the first and third stages is a difficult task in view that the propagation of the crack occurs at a seemingly instantaneous rate. On the other hand, the propagation in the second stage dominates the lifespan of the phenomenon of fatigue. Back in 1961, P. C. Paris and coworkers proposed that the crack growth rate in region II is governed by a power law [3]. Their propagation model is the most widely accepted nowadays, and it may be described as the initial-value problem

$$\begin{aligned} &\frac{dX_t}{dt} = cX_t^m, \quad t > 0, \\ &\text{subject to } X_0 = a. \end{aligned} \tag{2.1}$$

Here, c and m are positive constants that depend on each material, the number a satisfies $0 < a \ll 1$, and X_t represents the maximum length of the crack at time t.

There are many generalization of Paris' law reported in the literature [4]. Some of these extensions are deterministic models. However, some random components have been incorporated in Paris' law in order to account for unpredictable variability [5]. In particular, the work [6] investigates the stochastic model

$$dX_t = c|X_t|^m + \tilde{c}|X_t|^n dW_t, \quad t > 0, \tag{2.2}$$

where $\tilde{c}$ is a nonzero constant, c is a real number, W is a Brownian motion, and $m, n \in [1, \infty)$. The use of Itô's stochastic integral to interpret this differential equation results in the identity

$$X_t = a + c\int_0^t |X_s|^m ds + \tilde{c}\int_0^t |X_s|^n dW_s, \quad t > 0, \tag{2.3}$$

for some $a \in \mathbb{R}$. In this case, the explosion time actually corresponds to the time when the fracture occurs.

Assume that $m, n \in [1, \infty)$, and note that the continuous differentiability of $b(x) = cx^m$ and $\sigma(x) = \tilde{c}x^n$ for $x \in [0, \infty)$ assures the existence and the uniqueness of a solution of (2.3). Moreover, Proposition 1.1 guarantees that $X_t^a > 0$ is satisfied for $t > 0$ almost surely, whenever $a > 0$.

Proposition 2.1. *The solution of* (2.3) *explodes in finite time if* $\tilde{c} > 0$, $n > 1$ *and* $m > 2n - 1$.

In the sequel, we will use $\xi > 0$ to represent physically the maximal length of a crack. The next result identifies the distribution of the crack time τ_ξ^x.

Proposition 2.2. *Let* $c > 0$, *and suppose that* $m, n \in (0, \infty)$. *If* $u : [0, \infty) \times [0, \infty) \to \mathbb{R}$ *is a bounded function that satisfies the boundary-value problem*

$$\begin{aligned} &\frac{\partial u}{\partial t}(t,x) = \frac{\tilde{c}^2}{2}x^{2n}\frac{\partial^2 u}{\partial x^2}(t,x) + cx^m\frac{\partial u}{\partial x}(t,x), \quad (t,x) \in (0,\infty)\times(0,\xi), \\ &\text{subject to } \begin{cases} u(0,x) = 0, & x \in (0,\xi), \\ u(t,\xi) = 1, & t \geq 0, \end{cases} \end{aligned} \tag{2.4}$$

 ISBN: 978-84-608-6082-2

then $u(t,x) = P[\tau_\xi^x \leq t]$ *for each* $0 < x < \xi$ *and* $t > 0$. □

3 Aims of this work

The present work is devoted to designing a finite-difference discretization of a generalization of the problem studied in Theorem 1.2. Concretely, let ξ be a positive number and suppose that $b, \sigma : [0,\xi] \to \mathbb{R}$ are continuously differentiable functions such that $b(0) = \sigma(0) = 0$, and assume that $\varphi : (0,\xi) \to [0,1]$ is a nondecreasing function. For the remainder of this work, we will suppose that u is a real function defined for each $(t,x) \in [0,\infty) \times [0,\xi]$ satisfying the problem

$$\frac{\partial u}{\partial t}(t,x) - \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2}(t,x) - b(x)\frac{\partial u}{\partial x}(t,x) = 0, \quad (t,x) \in (0,\infty) \times (0,\xi),$$
$$\begin{cases} u(0,x) = \varphi(x), & x \in (0,\xi), \\ u(t,0) = 0, & t \geq 0, \\ u(t,\xi) = 1, & t \geq 0. \end{cases} \tag{3.1}$$

Obviously, in the practice we will consider $\varphi(x) = 0$ for each $x \in (0,\xi)$. Meanwhile, the functions b and σ will be power laws, so that the conditions $b(0) = \sigma(0) = 0$ will be readily satisfied.

References

[1] Henry P McKean. *Stochastic Integrals*. American Mathematical Society, New York, first edition, 1969.

[2] M. F. Kaplan. Crack propagation and the fracture of concrete. In *ACI Journal Proceedings*, volume 58. ACI, 1961.

[3] Paul C Paris, Mario P Gomez, and William E Anderson. A rational analytic theory of fatigue. *The Trend in Engineering*, 13(1):9–14, 1961.

[4] Nicola Pugno, M Ciavarella, Pietro Cornetti, and Alberto Carpinteri. A generalized Paris law for fatigue crack growth. *Journal of the Mechanics and Physics of Solids*, 54(7):1333–1349, 2006.

[5] J. Maljaars, H. M. G. M. Steenbergen, and A. C. W. M. Vrouwenvelder. Probabilistic model for fatigue crack growth and fracture of welded joints in civil engineering structures. *International Journal of Fatigue*, 38:108–117, 2012.

[6] Kazimierz Sobczyk. *Random Fatigue: From Data to Theory*. Kluwer Academic Publishers, Norwell, MA, United States of America, 1992.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Relationship of the conflation in the Belnap's logic with the (crisp) stable model semantics.

Nicolás Madrid[1]

[1] *Departamento de matemática aplicada, Universidad de Málaga*

emails: nicolas.madrid@uma.es

Abstract

In this paper I study the (crisp) stable model semantics by embedding crisp normal logic programs into the Belnap's fourth valued logic structure. The advantage of this consideration is that we can guarantee the existence of fuzzy stable models on such multi-valued logic structure. Specifically, I present some novel results relating fuzzy stable models of crisp logic programs and the conflation, an operator similar to negation defined on bilattice structures.

Key words: Logic programming, stable models, well-founded semantics, inconsistence, fuzzy logic programming, fuzzy answer set semantics, Belnap's Logic.

1 Introduction

Logic Programming[9] is a paradigm that combines the formality of logic and the efficiency of programming. From a logical point of view, it has evolved from a very simplified crisp logic to a family of complex theories like Abductive logic programming, Answer set programming, Constrain logic programming and Inductive logic programming among others.

In this paper we focus on the stable model semantics, the seed of the the Answer set programming, which models the *default negation* (also called *negation as failure*). The use of such a negation, instead of the ordinary one, is motivated by a programming point of view and by the so call *"close world assumption"* [19, 9]. The consideration of the default negation leads to two different semantics in logic programming namely, the stable model semantics [6, 7]and the well-founded semantics[5] (both defined at the early 90's). These two theories where related in [4, 16] by extending the stable model semantics to the fourth valued Belnap logic [1] and showing that the well-founded semantics coincides with one fuzzy stable model.

 ISBN: 978-84-608-6082-2

Despite the relationship between multivalued logic and the (crisp) stable model semantics given above, the theoretical interest of the community stopped for searching relationships between crisp stable model semantics and fuzzy logic programming. Oppositely, the community focus on the generalisation of the stable model semantics and the well-founded semantics in the fuzzy framework. In such respect, we remark a family of interesting extensions of the stable model semantics [2, 8, 10, 12, 13, 14, 17, 18] and of the well founded semantics [11]. This paper returns to the approach of Melvin Fitting and studies the relationship between the (crisp) stable model semantics and the set of fuzzy stable models under the fourth valued Belnap's logic. The structure of the set of stable models is initially studied in [3, 4] by means of the fix points of the immediate consequence operator. Nevertheless, in this paper we focus on the conflation (an operator typical from bilatices[3]) instead of on the immediate consequence operator. Thus, we show that there is a strong relationship between the conflation and the fuzzy stable models (in the Belnap's logic framework) obtained from crisp normal logic programs.

In this paper we begin by recalling the fuzzy stable model semantics based on arbitrary residuated lattices. Then, in Section 3 we restrict our study to the Belnap's fourth valued logic. Finally, in Section 4 we present some future work.

2 Fuzzy Stable Models

As stated above, the paradigm of Fuzzy Logic programming is based on the mathematical structure of residuated lattices.

Definition 1 *A* residuated lattice *is a triple* $\mathcal{L} = ((L, \leq), *, \leftarrow)$ *such that:*

1. $(L, \leq)$ *is a complete and bounded lattice with largest element 1 and least element 0.*
2. $(L, *, 1)$ *is a commutative monoid unit element 1.*
3. $*$ *and* $\leftarrow$ *form an adjoint pair, i.e:*

$$z \leq (x \leftarrow y) \text{ iff } y * z \leq x \quad \text{for all } x, y, z \in L.$$

In residuated lattices, the operator $*$ is seen as a conjunction and the operator $\leftarrow$ as an implication. In the rest of the paper we will consider a residuated lattice enriched with a negation operator, $(L, *, \leftarrow, \neg)$. The negation will model the notion of default negation often used in logic programming. As usual, a negation operator, over L, is any decreasing mapping $n\colon L \to L$ satisfying $n(0) = 1$ and $n(1) = 0$. It is convenient to point out that $\neg$ is semantically different from the ordinary negation used in fuzzy logic.

The syntax of normal and definite residuated logic programs assumes a set Π of propositional symbols.

Definition 2 *Given a residuated lattice with negation* $(L, *, \leftarrow, \neg)$*, a* normal residuated logic program $\mathbb{P}$ *is a set of weighted rules of the form*

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad \vartheta \rangle$$

where ϑ *is an element of* L *and* $p, p_1, \ldots, p_n$ *are propositional symbols.*

A normal residuated logic program $\mathbb{P}$ is said to be *definite* if it does not contain any negation operator. The semantics is given by the following definition.

Definition 3 *A fuzzy* L-interpretation *is a mapping* $I \colon \Pi \to L$*; that is, an* L*-fuzzy subset of propositional symbols. We say that* I satisfies *a rule* $\langle p \leftarrow \mathcal{B}; \quad \vartheta \rangle$ *if and only if* $I(\mathcal{B}) * \vartheta \leq I(p)$ *or, equivalently,* $\vartheta \leq I(p \leftarrow \mathcal{B})$*. Finally,* I *is a* model *of* $\mathbb{P}$ *if it satisfies all rules (and facts) in* $\mathbb{P}$*.*

It is worth to point out that every definite residuated logic program has a least model which can be reached by iterating the immediate consequence operator [10]. Such a model is called the least Herbrand model. Now we describe the adaptation of the approach given in [6] and [7] to normal residuated logic programs defined above. Let us consider a normal residuated logic program $\mathbb{P}$ together with a fuzzy L-interpretation I. To begin with, we will construct a new normal program $\mathbb{P}_I$ by substituting each rule in $\mathbb{P}$ such as

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad \vartheta \rangle$$

by the rule[1]

$$\langle p \leftarrow p_1 * \cdots * p_m; \quad \neg I(p_{m+1}) * \cdots * \neg I(p_n) * \vartheta \rangle$$

Notice that the new program $\mathbb{P}_I$ is definite, that is, does not contain default negation.

Definition 4 *The program* $\mathbb{P}_I$ *is called the* reduct *of* $\mathbb{P}$ *wrt the interpretation* I*.*

Definition 5 *Let* $\mathbb{P}$ *be a normal residuated logic program and let* I *be a fuzzy* L*-interpretation;* I *is said to be a* fuzzy stable model *of* $\mathbb{P}$ *iff* I *is the least model of* $\mathbb{P}_I$*.*

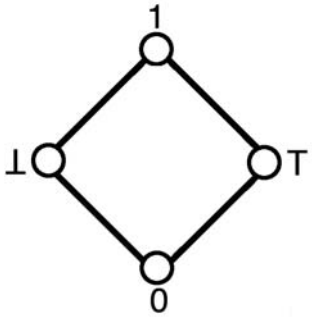
Theorem 1 *Any fuzzy stable model of* $\mathbb{P}$ *is a minimal model of* $\mathbb{P}$*.*

[1] Note the overloaded use of the negation symbol, as a syntactic function in the formulas and as the algebraic negation in the truth-values.

 ISBN: 978-84-608-6082-2

3 The specific case of the Belnap's fourth valued logic.

In this section we restrict our study to a very specific framework: the Belnap fourth valued logic. The lattice $\mathcal{B}$ that determines the set of truth values is given by Hasse diagram below

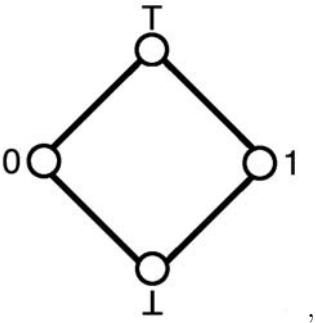


Moreover, the operators $*$ and $\neg$ are given by the following tables:

$*$	0	1	$\bot$	$\top$
0	0	0	0	0
1	0	1	$\bot$	$\top$
$\bot$	0	$\bot$	$\bot$	0
$\top$	0	$\top$	0	$\top$

$\neg$	0
0	1
1	0
$\bot$	$\bot$
$\top$	$\top$

Note that in residuated lattices the operator $\leftarrow$ is uniquely determined by $*$, so it is not necessary to be specified. The values 0 and 1 in $\mathcal{B}$ are interpreted as usual; i.e. they mean *true* and *false*, respectively. The other two values, $\bot$ and $\top$, mean *unknown* and *inconsistent*, respectively. One of the main features of the Belnap's logic connectives $*$ and $\neg$ is that they are monotonic with respect to a different order (in $\mathcal{B}$) from the one given above. Specifically with the order given by the following Hasse diagram



,

which is called *knowledge order*, the negation turns monotonic. Moreover, the operators $*$ and $\sup(\cdot)$[2] are monotonic as well. Then, it is possible to define a least model semantics for

[2]This operator supremum denotes the one with respect to the original order.

every normal residuated logic program by means of the immediate consequence operator, as it is done in general for definite logic programs. Moreover, it can be proved that such a model is in fact a fuzzy stable model as well. Thus, we have the following result.

Theorem 2 ([3]) *Every normal residuated logic program $\mathbb{P}$ defined on the Belnap fourth valued logic, has fuzzy stable models.*

What is interesting now is that every (crisp) normal logic program $\mathbb{P}$ can be seen as a logic program in $\mathcal{B}$ just by rewriting every rule

$$p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n$$

in $\mathbb{P}$ by

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad 1 \rangle$$

and by adding the rules

$$\langle q \leftarrow; \quad 0 \rangle$$

for any propositional symbol $q \in \Pi$ that does not appear in the head of any rule in $\mathbb{P}$. By the Theorem 2, $\mathbb{P}$ has at least one stable model in $\mathcal{B}$. Moreover, the least stable model under the knowledge ordering, which always exists, coincides with the Well-founded semantics [4]. Note that every (crisp) stable model is also a fuzzy stable model in $\mathcal{B}$, so we have the following result.

Proposition 1 *A (crisp) normal logic program $\mathbb{P}$ has no (crisp) stable models, if for every fuzzy stable model M of $\mathbb{P}$ in $\mathcal{B}$, there exists a propositional symbol $p \in \Pi$ such that $M(p) \notin \{0, 1\}$.*

The rest of the section presents a set of novel results with the aim of providing a structure in the set of stable models in $\mathcal{B}$. The first result shows that every stable model of $\mathbb{P}$ has another stable model associated by conflation. Let us recall that the *conflation* in $\mathcal{B}$ is an operator defined by:

$-$	
0	0
1	1
$\bot$	$\top$
$\top$	$\bot$

Note that the conflation is like a negation but, instead of reversing the natural order, it reverses only the knowledge order. Given a $\mathcal{B}$-fuzzy interpretation I, we define $-I = \{-I(p) \mid p \in \Pi\}$. Similarly we can apply the operator $-$ to programs as follows. Let $\mathbb{P}$ a normal residuated logic program. Then $-\mathbb{P}$ is defined by changing every rule

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad \vartheta \rangle$$

 ISBN: 978-84-608-6082-2

of $\mathbb{P}$ by

$$\langle p \leftarrow p_1 * \cdots * p_m * \neg p_{m+1} * \cdots * \neg p_n; \quad -\vartheta \rangle$$

Note that $-\mathbb{P}$ only modifies the weights of rules. We have the following result.

Proposition 2 *Let $\mathbb{P}$ be a definite residuated logic program and let $M_{\mathbb{P}}$ be the least model of $\mathbb{P}$ (under the original order). Then, $-M_{\mathbb{P}}$ is the least model of $-\mathbb{P}$.*

In the case of starting from a crisp normal logic program, the proposition above allows us to reach the following result:

Theorem 3 *Let $\mathbb{P}$ be a (crisp) normal residuated logic program and let M be a fuzzy stable model of $\mathbb{P}$ in $\mathcal{B}$. Then, $-M$ is also a fuzzy stable model.*

Note that if M is a crisp stable model, i.e. $M(p) \in \{0, 1\}$, then $-M = M$, and the result above does not define any new fuzzy stable model. In the case of an inconsistent crisp normal logic program, the theorem above implies the following result:

Corollary 1 *Let $\mathbb{P}$ be a (crisp) normal residuated logic program without (crisp) stable models. Then $\mathbb{P}$ has an even number of fuzzy stable models in $\mathcal{B}$.*

4 Conclusion and Future Work

In this paper we have shown that the conflation in the Belnap's fourth valued logic has convenient properties with the stable model semantics. Specifically we have shown that given a fuzzy stable model, we can define another by means of $-$. I think this result is a good starting point to relate the inconsistency in stable model semantics with fuzzy stable models in the Belnap's Logic. In such line, note that for every fuzzy stable model M there is at least a propositional symbol $p \in \Pi$ such that $M(p) \notin \{0, 1\}$. Then Theorem 3 can be applied to define new fuzzy stable models and, probably to relate cycle causing the inconsistency. On the same line, perhaps it is convenient to use different results about the existence of fuzzy stable models (as the one given in [15]) to determine the causes of inconsistence of crisp logic programs.

References

[1] N. D. Belnap. A useful four-valued logic. In G. Epstein and J. M. Dunn, editors, *Modern Uses of Multiple-Valued Logic*, pages 7–37. Reidel Publishing Company, Boston, 1977.

[2] C. V. Damásio and L. M. Pereira. Antitonic logic programs. In *Proc. Logic Programming and Nonmonotonic Reasoning, LPNMR'01*, pages 379–392. Lect. Notes in Artificial Intelligence, 2173, Springer-Verlag, 2001.

[3] M. Fitting. Bilattices and the semantics of logic programming. *Journal of Logic Programming*, 11:91–116, 1991.

[4] M. Fitting. The family of stable models. *The Journal of Logic Programming*, 17(2-4):197 – 225, 1993.

[5] A. V. Gelder, K. A. Ross, and J. S. Schilpf. The well-founded semantic for general logic programs. *Journal of the ACM*, 38(3):620–650, July 1991.

[6] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In *Proc. of ICLP-88*, pages 1070–1080, 1988.

[7] M. Gelfond and V. Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9:365–385, 1991.

[8] J. Janssen, S. Schockaert, D. Vermeir, and M. De Cock. General fuzzy answer set programs. *Lecture Notes in Computer Science*, 5571(352–359), 2009.

[9] J. Lloyd. *Foundations of Logic Programming*. Springer Verlag, 1987.

[10] Y. Loyer and U. Straccia. Epistemic foundation of stable model semantics. *Journal of Theory and Practice of Logic Programming*, 6:355–393, 2006.

[11] Y. Loyer and U. Straccia. The Approximate Well-founded Semantics for Logic Programs with Uncertainty. *Lecture Notes in Computer Science*, 3153:513-524, 2004.

[12] T. Lukasiewicz. Fuzzy description logic programs under the answer set semantics for the semantic web. *Fundamenta Informaticae*, 82(3):289–310, 2008.

[13] N. Madrid and M. Ojeda-Aciego. Measuring inconsistency in fuzzy answer set semantics. *IEEE Transactions on Fuzzy Systems*, 2011. vol. 19, no. 4, pp. 605–622, 2011.

[14] N. Madrid and M. Ojeda-Aciego. On the existence of stable models in normal residuated logic programs. In *Computational Methods in Mathematics, Science, and Engineering*, pages 598–604, 2010.

[15] N. Madrid and M. Ojeda-Aciego. On the use of fuzzy stable models for inconsistent classical logic programs. In *IEEE Symposium on Foundations of Computational Intelligence, FOCI 2011*.

[16] T. Przymusinski. Well-founded semantics coincides with three-valued stable semantics. *Fundamenta Informaticae*, 13:445–463, 1990.

[17] U. Straccia. Annotated answer set programming. Technical Report 2005-TR-51, Istituto di Scienza e Tecnologie dell'Informazione-Consiglio Nazionale delle Ricerche, 2005.

 ISBN: 978-84-608-6082-2

[18] D. Van Nieuwenborgh, M. De Cock, and D. Vermeir. An introduction to fuzzy answer set programming. *Ann. Math. Artif. Intell.*, 50(3-4):363–388, 2007.

[19] G. Wagner. Web rules need two kinds of negation. *Lecture Notes in Computer Science*, 2901:33–50, 2003.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Improved convergence analysis for Newton-like methods

Á. Alberto Magreñán[1], Ioannis K. Argyros[2] and Juan Antonio Sicilia[1]

[1] *Escuela de Ingeniería, Universidad Internacional de La Rioja (UNIR)*

[2] *Department of Mathematics Sciences, Cameron University*

emails: alberto.magrenan@unir.net, iargyros@cameron.edu,
juanantonio.sicilia@unir.net

Abstract

We present a new semilocal convergence analysis for Newton-like methods in order to approximate a locally unique solution of an equation in a Banach space setting. This way, we expand the applicability of these methods in cases not covered in earlier studies. The new analysis is based on our new idea of restricted convergence domains. Using this idea we determine a more precise location, where the iterates lie leading to tighter Lipschitz constants. The advantages of our approach include a more precise convergence analysis under the same computational cost on the Lipschitz constants involved.

Key words: Newton-type method, Banach space, majorizing sequence, divided difference, semilocal-local convergence.
MSC 2000: 65H10, 65G99, 65B05,65N30, 47J25, 47J05

1 Introduction

Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces. Let $U(w,R)$ and $\overline{U}(w,R)$ stand, respectively, for the open and closed ball centered at $w \in \mathcal{X}$ and of radius $R > 0$. Let also $\text{Ł}(\mathcal{X},\mathcal{Y})$ denote the space of bounded linear operators from $\mathcal{X}$ into $\mathcal{Y}$.

In this study we are concerned with the problem of approximating a locally unique solution of the equation

$$F(x) = 0, \tag{1.1}$$

where F is a nonlinear operator defined on a convex subset $\mathcal{D}$ of $\mathcal{X}$ with values in $\mathcal{Y}$. Many Problems from Computational Sciences and other disciplines can be brought in the form of an equation like (1.1) using Mathematical Modelling [2, 3, 4]. The solutions of theses equations can be given in closed form only in special cased. That is why most solution methods for these equations are usually iterative.

A lot of iterative methods for solving equation (1.1) can be written as Newton-like methods in the form

$$x_{n+1} = x_n - T_n F(x_n) \text{ for each } n = 0, 1, 2, \ldots, \tag{1.2}$$

where $T_n \in \text{Ł}(\mathcal{Y},\mathcal{X})$ and $x_0 \in \mathcal{X}$ is an initial point.

 ISBN: 978-84-608-6082-2

Potra [8, 9] studied the semilocal convergence of iterative algorithm (1.2) when

$$T_n \in \{\delta F(x_{p_n}, x_{q_n})^{-1}, \delta F(x_{q_n}, x_{p_n})^{-1}\} \quad \text{for each } n = 0, 1, 2, \ldots, \tag{1.3}$$

where $\{p_n\}$ and $\{q_n\}$ are two nondecreasing sequences of integers satisfying

$$q_0 = -1, \quad p_0 = 0, \quad q_n \leq p_n \leq n \quad \text{for each } n = 0, 1, 2, \ldots. \tag{1.4}$$

Potra provided under a popular sufficient convergence criterion (to be precised in (2.5)) optimal (in some sense) error estimates on the distances $\|x_n - x_{n+1}\|$ and $\|x^* - x_n\|$ for each $n = 0, 1, 2, \ldots$.

In the present study, we are motivated by Potra's work in [8], related studies [1, 5, 6, 7, 10] and optimization considerations. Using center-Lipschitz condition (see (2.9)) instead of the less precise Lipschitz condition (see (2.8)), we obtain more precise estimates on the upper bounds of $\|T_n\|$ leading to more precise majorizing sequences for $\{x_n\}$. The advantages of our approach are:

(a) More precise error estimates on the distances $\|x_n - x_{n+1}\|$ and $\|x^* - x_n\|$ for each $n = 0, 1, 2, \ldots$;

(b) At least as precise information on the location of the solution.

These advantages are obtained under the same computational cost on the Lipschitz constants involved as in the earlier studies.

The rest of the paper is organized as follows: The semilocal convergence analysis is presented in Section 2 and the local analysis in Section 3.

2 Semilocal Convergence

In this Section we present the semilocal convergence analysis of iterative procedure (1.1)-(1.2) for triplets (F, x_{-1}, x_0) belonging to the class $C(k, k_0, k_1, k_2, b, c)$ defined as follows.

Definition 2.1 *Let $k > 0$, $k_0 > 0$, $k_1 > 0$, $k_2 \geq 0$, $b \geq 0$ and $c \geq 0$. Suppose that*

$$k^* b + 2\sqrt{k^* c} \leq 1, \tag{2.5}$$

where $k^* = max\{k, k_0, k_1, k_2\}$.

We say that a triplet $F(F, x_0, x_{-1})$ belongs to the class $C(k, k_0, k_1, k_2, b, c)$, if:

(C_1) *F is a nonlinear operator defined on a convex subset $\mathcal{D}$ of a Banach space $\mathcal{X}$ with values in a Banach space $\mathcal{Y}$.*

(C_2) *x_0 and x_{-1} are two points belonging to the interior $\overset{\circ}{\mathcal{D}}$ of $\mathcal{D}$ and satisfying the inequality*

$$\|x_0 - x_{-1}\| \leq b; \tag{2.6}$$

(C_3) *F is Fréchet differentiable on $\overset{\circ}{\mathcal{D}}$ and there exist a mapping $\delta F : \overset{\circ}{\mathcal{D}} \times \overset{\circ}{\mathcal{D}} \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$ such that the linear operator P_0, where P_0 is either $\delta F(x_0, x_{-1})$ or $\delta F(x_{-1}, x_0)$, is invertible, its inverse $T_0 = P_0^{-1}$ is bounded and constants $k > 0$, $k_0 > 0$, $k_1 > 0$, $k_2 \geq 0$, $b \geq 0$ and $c \geq 0$ are such that:*

$$\|T_0 F(x_0)\| \leq c; \tag{2.7}$$

(C_4)
$$\|T_0(\delta F(x,y)-F'(x_0))\| \leq k_0\|x-x_0\|+k\|y-x_0\| \text{ for each } x,y\in\mathcal{D} \tag{2.8}$$
and
$$\|T_0(\delta F(x,y)-F'(z))\| \leq k_1\|x-z\|+k_2\|y-z\| \text{ for each } x,y,z\in\mathcal{D}\cap U(x_0,\frac{1}{k_0+k})=\mathcal{D}_1; \tag{2.9}$$

(C_5) *The set* $\mathcal{D}_c=\{x\in\mathcal{D}; F \text{ is continuous at } x\}$ *contains the closed ball* $\bar{U}(x_0,t^*-b)$ *or the ball* $U(x_0,\frac{1}{k_0+k})$, *where*
$$t^*=\frac{1+k^*b-\sqrt{(1+k^*b)^2-4k^*(b+c)}}{2k^*}.$$

Next, we present the main semilocal convergence result for the Newton-like method (1.2) using the notation introduced above.

Theorem 2.2 *Suppose that* $(F,x_{-1},x_0)\in C(k,k_0,k_1,k_2,b,c)$ *and* $T_n=\delta F(x_{p_n},x_{q_n})^{-1}$. *Then, sequence* $\{x_n\}$ $(n\geq -1)$ *generated by the Newton-like method* (1.2) *is well defined, remains in* $\bar{U}(x_0,t^*-b)$ *for each* $n=0,1,2,\ldots$ *and converges to a solution* $x^*\in\bar{U}(x_0,t^*-b)$ *of equation* $F(x)=0$.

Concerning the existence and the uniqueness of the solution for nonlinear equations, we present the following result:

Corollary 2.3 *Suppose that* $(F,x_{-1},x_0)\in C(k,k_0,k_1,k_2,b,c)$. *Then, the equation* $F(x)=0$ *has a solution* $x^*\in D$ *and this is the unique solution of the equation in the set* $V=\{x\in D:\|x-x_0\|<\bar{t}_0+d\}$ *if* $t_0>0$ *or in the set* $W=\{x\in D:\|x-x_0\|<\bar{t}_0\}$ *if* $d=0$, *where* $\bar{t}_0=\frac{1+k_1b}{2k_1}$ *and*
$$d=\frac{\sqrt{(1+k^*b)^2-4k^*(b+c)}}{2k^*}.$$

3 Local Convergence

We study the local convergence analysis for the Newton-like method (1.2) for couples (F,x^*) belonging to the class $A(l,l_0,l_1,l_2)$ if:

(A_1) F is a nonlinear operator defined on a convex subset $\mathcal{D}$ of a Banach space $\mathcal{X}$ with values in a Banach space $\mathcal{Y}$.

(A_2) F is Fréchet differentiable on $\overset{\circ}{\mathcal{D}}$ and there exist a mapping $\delta F:\overset{\circ}{\mathcal{D}}\times\overset{\circ}{\mathcal{D}}\to Ł(\mathcal{X},\mathcal{Y})$ and constants $l>0$, $l_0>0$, $l_1>0$ and $l_2\geq 0$ such that:
$$F(x^*)=0,\quad F'(x^*)^{-1}\in Ł(\mathcal{Y},\mathcal{X}),$$
$$\|F'(x^*)^{-1}(\delta F(x,y)-F'(x^*))\|\leq l_0\|x-x^*\|+l\|y-x^*\| \text{ for each } x,y\in\mathcal{D}$$
and
$$\|F'(x^*)^{-1}(\delta F(x,y)-F'(z))\|\leq l_1\|x-z\|+l_2\|y-z\| \text{ for each } x,y,z\in\mathcal{D}\cap U(x^*,\frac{1}{l_0+l})$$

(A_3) The set $\mathcal{D}_c=\{x\in\mathcal{D}; F \text{ is continuous at } x\}$ contains the ball $\bar{U}(x^*,R)$ or the ball $U(x^*,\frac{1}{l_0+l})$, where
$$R=\frac{2}{5(l_1+l_2)+l_0+l}.$$

 ISBN: 978-84-608-6082-2

Using the preceding notation we can show the main local convergence result for Newton-like method (1.2).

Theorem 3.1 *Suppose that* $(F, x^*) \in A(l, l_0, l_1, l_2)$ *and* $T_n = \delta F(x_{p_n}, x_{q_n})^{-1}$. *Then sequence* $\{x_n\}$ $(n \geq -1)$ *generated by Newton-like method* (1.2) *for* $x_{-1}, x_0 \in U(x^*, R)$ *is well defined, remains in* $U(x_0, R)$ *for each* $n = -1, 0, 1, 2, \ldots$ *and converges to* x^*. *Moreover, the following estimates hold*

$$\|x_{n+1} - x^*\| \leq \frac{\left[\frac{l_1 + l_2}{2}\|x_n - x^*\| + l_1\|x_n - x_{p_n}\| + l_2\|x_n - x_{q_n}\|\right]\|x_n - x^*\|}{1 - (l_0\|x_{p_n} - x^*\| + l\|x_{q_n} - x^*\|)} < R, \qquad (3.10)$$

Acknowledgements

This research was supported by Universidad Internacional de La Rioja (UNIR, http://www.unir.net), under the Plan Propio de Investigación, Desarrollo e Innovación 3 [2015–2017]. Research group: MOdelación Matemática Aplicada a la INgeniería(MOMAIN), by the the grant SENECA 19374/PI/14 and by Ministerio de Ciencia y Tecnología MTM2014-52016-C2-01-P

Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] Amat, S., Busquier, S., Negra, M., Adaptive approximation of nonlinear operators, Numer. Funct. Anal. Optim., 25 , 397–405 (2004).

[2] Argyros, I.K., Computational theory of iterative methods. Series: Studies in Computational Mathematics, 15, Editors: C.K. Chui and L. Wuytack, Elsevier Publ. Co. New York, U.S.A. (2007).

[3] Argyros, I.K., Hilout, S., Weaker conditions for the convergence of Newton's method, Journal of Complexity, AMS, 28, 364–387 (2012).

[4] I. K. Argyros and Y. J. Cho, S. Hilout, Numerical method for equations and its applications. CRC Press/Taylor and Francis, New York (2012).

[5] Bosarge, W.E. and Falb, P.L., A multipoint method of third order, J. Optimiz. Th. Appl., 4, 156-166 (1969).

[6] Bosarge, W.E., Falb, P.L., Infinite dimensional multipoint methods and the solution of two point boundary value problems, Numer. Math., 14, 264-286 (1970).

[7] Cătinaş, E., The inexact, inexact perturbed, and quasi-Newton methods are equivalent models, Math. Comp., 74 (249), 291–301 (2005).

[8] F.A. Potra, Sharp error bounds for a class of Newton-like methods, Libertas Math., 5, 71–84 (1985).

[9] Potra, F.A., Pták, V., Nondiscrete induction and iterative processes. Research Notes in Mathematics, 103. Pitman (Advanced Publishing Program), Boston, MA (1984).

[10] Magreñán, A.A., Argyros, I.K., Improved convergence analysis for Newton-like methods, Numerical Algorithms, 1–23 (2015).

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Some novel and optimal families of King's method with eighth and sixteenth-order of convergence

P. Maroju[1], R. Behl[1] and S.S. Motsa[1]

[1] *School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal, Private Bag X01, Scottsville 3209, Pietermaritzburg, South Africa,*

emails: marojuprashanth@gmail.com, ramanbehl87@yahoo.in, sandilemotsa@gmail.com

Abstract

In this study, our principle aim is to provide some novel eighth and sixteenth-order families of King's method for solving nonlinear equations which should be superior than the existing schemes of same order. The relevant optimal orders of the proposed families satisfy the classical Kung-Traub conjecture which was made in 1974. The derivation of the proposed schemes is based on the weight function and rational approximation approaches. In addition, convergence properties of the proposed families are fully investigated along with two main theorems and one lemma describing their order of convergence. We consider a concrete variety of numerical examples to check the validity and effectiveness of the our proposed methods. Further, it is found from the numerical results that our proposed methods perform better than the existing ones of the same order when the accuracy is checked in the multi precision digits by choosing the same numerical examples with same initial guesses.

Key words: Order of convergence, Newton's method, Kung-Traub conjecture.
MSC 2000: AMS codes (optional)

1 Introduction

Finding the solution techniques to solve the nonlinear equations, have always been a paramount importance in the field of numerical analysis which provide the accurate and efficient approximate solution α of a nonlinear equation of the form

$$f(x) = 0, \tag{1}$$

 ISBN: 978-84-608-6082-2

where $f : I \subseteq \mathbb{R} \to \mathbb{R}$ is a sufficiently differentiable function on the interval I.

One of the main reason of paramount importance of this topic is the applicability in the applied science and the four major disciples of engineering: chemical, electrical, civil and mechanical (for the detailed explanation please the Chapra and Canale [11]). For example, the location of the extremal points of a function describing some system requires finding the zeros of the derivatives of that function. Many problems which involve critical paths also require the solution of algebraic equations, such as determining all the ray paths that are possible in a complex optical system and can be modeled by different mathematical equations.

Newton's method [1–26] is one of the most basic and popular iterative methods for solving nonlinear equations. But, it is a one-point iterative method and one-point methods have some drawbacks regarding their efficiency and convergence order. Therefore, several scholars from the worldwide turned towards multi-point iterative methods. The advantage of multi-point methods is that they do not use higher order derivatives and has great practical importance because they overcome from the theoretical limitations of one-point methods (for the details please see [25]).

Few year ago, it was not an easy task to propose higher-order multi-point iterative methods. But, with the advancement of digital computer, advanced computer arithmetics and symbolic computation, the construction of higher-order multi-point methods become more vital and popular because they provide more accurate and efficient approximated root with in a very small number of iterations and their efficiency index [25] is better than the classical Newton's method. Therefore, in the last two decades, a variety of three-point optimal eighth-order multi-point methods, without memory have been proposed in [1–5, 12–15, 19, 21–24, 26]. Most of them are the extension of Newton's method or Newton like method at the expense of additional functional evaluations or sub steps.

In the past, Kung and Traub [18] given two general classes of optimal n-point iterative methods. Then later, Neta [9], presented an optimal family of sixteenth-order iterative methods. Recently, Li et al. [6], Guem and Kim [10, 15], Sharma et al. [7], Ullah et al. [8], have also proposed optimal sixteenth-order multi-point methods except the Li et al. [6].

In this paper, we proposed some new novel eighth-order family of King's method which is based on the weight function approach. Then, we will extend this family from eighth-order convergence to sixteenth-order with the aid of rational functional approximations approach. The proposed families of King's method have optimal order of convergence in the sense of Kung-Traub conjecture. The efficiency of the methods is tested on a concrete variety of test functions. From the numerical experiments, it is observed that our proposed methods perform better than existing methods of same order.

2 An optimal family of eighth-order King's method

In this section, we will present an optimal eighth-order family of King's methods. Therefore, we consider the following three-step scheme:

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - \frac{f(x_n)+\beta f(y_n)}{f(x_n)+(\beta-2)f(y_n)}\frac{f(y_n)}{f'(x_n)}, \quad \beta \in \mathbb{R}, \\ x_{n+1} &= z_n - \frac{f(z_n)}{f'(x_n)}G(u,\ v), \end{aligned} \tag{2}$$

where the above weight function $G : \mathbb{R}^2 \to \mathbb{R}$, is a sufficiently differentiable function in the neighborhood of $(0,\ 0)$ and

$$u = \frac{f(z_n)}{f(y_n)}, \quad v = \frac{f(y_n)}{f(x_n)}. \tag{3}$$

Since, the above scheme (2) uses only four functional evaluations and according to Kung-Traub conjecture its maximum order can be eight. In the following results, we will discuss the conditions on the weight function so that we will reach at an optimal eighth-order of convergence.

Lemma 2.1 *Let us assume that α be a simple zero of the involved function f. Further, we also assume that the function $f : I \subseteq \mathbb{R} \to \mathbb{R}$, is a sufficiently differentiable function defined in an interval containing the required zero α. Then, the quotients defined in* (3) *will satisfy the following error equations*

$$u = \frac{f(z_n)}{f(y_n)} = O(e_n^2), \quad v = \frac{f(y_n)}{f(x_n)} = O(e_n). \tag{4}$$

Proof Let us assume that $e_n = x_n - \alpha$, $e_{n,y} = y_n - \alpha$ and $e_{n,z} = z_n - \alpha$ be the errors in the n^{th} iteration. Now, we can expand the function $f(x_n)$ around the point $x = \alpha$ with the help of the Taylor's series expansion which will leads us to:

$$f(x_n) = f'(\alpha)\left(e_n + c_2e_n^2 + c_3e_n^3 + c_4e_n^4 + c_5e_n^5 + c_6e_n^6 + c_7e_n^7 + c_8e_n^8 + O(e_n^9)\right), \tag{5}$$

where $f'(\alpha) \neq 0$ and $c_k = \frac{1}{k!}\frac{f^k(\alpha)}{f'(\alpha)}$, $k = 2,\ 3,\ \ldots, 8$.
Similarly, we will obtain

$$f'(x_n) = f'(\alpha)\left(1 + 2c_2e_n + 3c_3e_n^2 + 4c_4e_n^3 + 5c_5e_n^4 + 6c_6e_n^5 + 7c_7e_n^6 + 8c_8e_n^7 + 9c_9e_n^8 + O(e_n^9)\right). \tag{6}$$

By inserting the equations (5) and (6) in the first sub step of the scheme (2), we have

$$e_{n,y} = c_2 e_n^2 + (2c_3 - 2c_2^2)e_n^3 + \sum_{j=1}^{5} A_j e_n^{j+3} + O(e_n^9), \tag{7}$$

where $A_j = A_j(c_2,\ c_3,\ \ldots\ c_8)$.
With the help of the above equation (7) and Taylor's series expansion, we will further obtain

$$f(y_n) = f'(\alpha)\left(e_{n,y} + c_2 e_{n,y}^2 + c_3 e_{n,y}^3 + c_4 e_{n,y}^4 + O(e_n^9)\right). \tag{8}$$

Further, with the help of previous equations (5) and (8), we will further yield

$$v = \frac{f(y_n)}{f(x_n)} = c_2 e_n + (2c_3 - 3c_2^2)e_n^2 + (8c_2^3 - 10c_3c_2 + 3c_4)e_n^3 + \sum_{j=1}^{5} D_j e_n^{j+3} + O(e_n^9), \tag{9}$$

where $D_j = D_j(c_2,\ c_3,\ \ldots\ c_8)$.
Now, by inserting the equations (5)–(8), in the second sub step of (2), we have

$$e_{n,z} = \left((2\beta+1)c_2^3 - c_2c_3\right)e_n^4 + \sum_{j=1}^{4} B_j e_n^{j+4} + O(e_n^9), \tag{10}$$

where $B_j = B_j(\beta,\ c_2,\ c_3,\ \ldots\ c_8)$.
In the similar fashion as we did in the previous equation (5), we can expand the function $f(z_n)$ about a point $x = \alpha$, which is given by

$$f(z_n) = f'(\alpha)\left(e_{n,z} + c_2 e_{n,z}^2 + O(e_n^9)\right). \tag{11}$$

From the equations (8) and (11), we have

$$u = \frac{f(z_n)}{f(y_n)} = \left((2\beta+1)c_2^2 - c_3\right)e_n^2 + \sum_{j=1}^{6} H_j e_n^{j+2} + O(e_n^9), \tag{12}$$

where $H_j = H_j(\beta,\ c_2,\ c_3,\ \ldots\ c_8)$.
This complete the proof of lemma 2.1. □

Theorem 2.2 *Let us assume that an initial guess $x = x_0$ is sufficiently close to α for the guaranteed convergence. Then, the iterative scheme* (2) *will reach an optimal eighth-order convergence only if it satisfies the following conditions on the weight function*

$$G_{00} = 1,\ G_{01} = 2,\ G_{10} = 1,\ G_{02} = 10 - 4\beta,\ G_{11} = 4,\ G_{03} = 12(\beta^2 - 6\beta + 6), \tag{13}$$

where $G_{ij} = \frac{\partial^{i+j}}{\partial u^i \partial v^j} G(u,\ v)|_{(u=0,\ v=0)}$ for $i, j = 0, 1, 2, 3$.

Proof Since, it is clear from the above lemma 2.1 that $u = O(e_n^2)$ and $v = O(e_n)$. Therefore, we can expand the weight function $G(u,\ v)$ in the neighborhood of $(0,\ 0)$ with the help Taylor series expansion which leads to us:

$$\begin{aligned} G(u,\ v) = G_{00} + G_{10}u + G_{01}v + \frac{1}{2!}\big(G_{20}u^2 + 2G_{11}uv + G_{02}v^2\big) \\ + \frac{1}{3!}\big(G_{30}u^3 + 3G_{21}u^2v + 3G_{12}uv^2 + G_{03}v^3\big). \end{aligned} \tag{14}$$

By using the equations (5)–(12) and (14), in the last sub step of the proposed scheme, we will get

$$e_{n+1} = -(G_{00}-1)c_2\left((2\beta+1)c_2^2 - c_3\right)e_n^4 + \sum_{j=1}^{4}\widetilde{H}_j e_n^{j+4} + O(e_n^9). \tag{15}$$

It is straightforward to say that the above error equation (15), will reach at least fifth-order of convergence if we choose the following value of G_{00}

$$G_{00} = 1. \tag{16}$$

Now, by inserting the above value of $G_{00} = 1$ in $\widetilde{H}_1 = 0$, we will further obtain

$$G_{01} = 2. \tag{17}$$

Again, by using the above values of G_{00} and G_{01} in $\widetilde{H}_2 = 0$, we will obtain the following two independent relations

$$G_{10} - 1 = 0, \quad G_{02} + 2(2\beta G_{10} + G_{10} - 6) = 0, \tag{18}$$

which further yield

$$G_{10} = 1, \qquad G_{02} = 10 - 4\beta. \tag{19}$$

In order to obtain an optimal eighth-order of convergence, we have to use the above values of $G_{00},\ G_{01}, G_{10}$ and G_{02} in $\widetilde{H}_3 = 0$. Then, we obtain

$$G_{11} - 4 = 0,\ \ G_{03} + 6\left(-2\beta^2 + 4\beta + 2\beta G_{11} + G_{11} - 16\right) = 0, \tag{20}$$

which will further leads to us

$$G_{11} = 4,\ G_{03} = 12(\beta^2 - 6\beta + 6). \tag{21}$$

Finally, we will obtain the following error equation by using the equations (16), (17), (19) and (21) in (15), which is given by

$$\begin{aligned} e_{n+1} = -\frac{c_2((2\beta+1)c_2^2 - c_3)}{2}\Big[c_2^4\left(4\beta^3 - 32\beta^2 + 44\beta + 2\beta G_{12} + G_{12} + (2\beta+1)^2 G_{20} - 82\right) \\ + c_3^2(G_{20} - 2) - 2c_4c_2 - c_3c_2^2(-4\beta + G_{12} + 4\beta G_{20} + 2G_{20} - 30)\Big]e_n^8 + O(e_n^9). \end{aligned} \tag{22}$$

 ISBN: 978-84-608-6082-2

The above error equation prove that our proposed scheme (2) reaches at eighth-order convergence by consuming only four functional evaluations per iteration. Further, the scheme (2) have also reached the optimal order of convergence in the sense of Kung-Traub conjecture. This completes the proof. □

2.1 Special cases

In this section, we will discuss some of the important cases of the weight function in two different approaches, both approaches are defined as follows:

(1) Let us consider the following weight function

$$G(u,v)=\frac{2\beta+u\left(2\beta+2(\beta^2-2\beta-4)v-5\right)-(4\beta+1)v^2+2(\beta^2-4\beta+1)v-5}{2\beta+2\left(\beta^2-6\beta+6\right)v-5}, \tag{23}$$

which will further yield

$$\begin{aligned} y_n &= x_n-\frac{f(x_n)}{f'(x_n)},\\ z_n &= y_n-\frac{f(x_n)+\beta f(y_n)}{f(x_n)+(\beta-2)f(y_n)}\frac{f(y_n)}{f'(x_n)},\\ x_{n+1} &= z_n-\frac{f(z_n)}{f'(x_n)}\left(\frac{2\beta+u\left(2\beta+2(\beta^2-2\beta-4)v-5\right)-(4\beta+1)v^2+2(\beta^2-4\beta+1)v-5}{2\beta+2\left(\beta^2-6\beta+6\right)v-5}\right). \end{aligned} \tag{24}$$

In this way, we obtain a new optimal eighth-order family of King's method.

(2) In order to obtain another new optimal eighth-order family of King's method, we consider the following weight function

$$G(u,v)=1+u+4uv-\frac{(4\beta+1)v}{2\left(\beta^2-6\beta+6\right)}+\frac{a_2v}{a_1v+1}, \tag{25}$$

where $a_1=\frac{2\left(\beta^2-6\beta+6\right)}{2\beta-5}$ and $a_2=\frac{2\beta-5}{a_1}$.
With the aid of the above weight function, we will further obtain

$$\begin{aligned} y_n &= x_n-\frac{f(x_n)}{f'(x_n)},\\ z_n &= y_n-\frac{f(x_n)+\beta f(y_n)}{f(x_n)+(\beta-2)f(y_n)}\frac{f(y_n)}{f'(x_n)},\\ x_{n+1} &= z_n-\frac{f(z_n)}{f'(x_n)}\left(1+4uv+u-\frac{(4\beta+1)v}{2\left(\beta^2-6\beta+6\right)}+\frac{a_2v}{a_1v+1}\right). \end{aligned} \tag{26}$$

(3) In this case, we will use the conditions on G_{ij} (which are defined in (13)) in the weight function (14). This is the another way of obtaining the new weight functions, which is given as follows:

$$\begin{aligned} G(u,v) = 1 + u + 2v + \frac{1}{2}\left(G_{20}u^2 + 8uv + (10-4\beta)v^2\right) \\ + \frac{1}{6}\left(3G_{12}uv^2 + 3G_{21}u^2v + G_{30}u^3 + 12\left(\beta^2 - 6\beta + 6\right)v^3\right), \end{aligned} \tag{27}$$

where $G_{20},\ G_{12},\ G_{21}$ and G_{30} are free disposable parameters. By using the above weight function and the values of the disposable parameters in the scheme (2), we will obtain several new optimal eighth-order families of King's method.

3 An optimal family of sixteenth-order King's method

In this section, we will propose a new optimal sixteenth-order family of iterative methods. The idea is consider one more step in the family (2). So if we use the same notation than in the previous section, we consider

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - \frac{f(x_n) + \beta f(y_n)}{f(x_n) + (\beta-2)f(y_n)} \frac{f(y_n)}{f'(x_n)}, \\ t_n &= z_n - \frac{f(z_n)}{f'(x_n)} G(u,\ v), \end{aligned} \tag{28}$$

In order to obtain the next approximation x_{n+1} to the required root, we consider

$$Q(x) = \frac{(x - x_n) + \theta_1}{\theta_2(x-x_n)^3 + \theta_3(x-x_n)^2 + \theta_4(x-x_n) + \theta_5}, \tag{29}$$

a rational function, where $\theta_1,\ \theta_2,\ \theta_3,\ \theta_4$ and θ_5 are disposable parameters. Further, these parameters can be determined by imposing tangency conditions, which are given by

$$Q(x_n) = f(x_n),\ \ Q'(x_n) = f'(x_n),\ \ Q(y_n) = f(y_n),\ \ Q(z_n) = f(z_n),\ \ Q(t_n) = f(t_n). \tag{30}$$

Now, we assume that the above rational function meets the x – axis at $x = x_{n+1}$ in order to find the next approximation, which is given by

$$Q(x_{n+1}) = 0, \tag{31}$$

which further yields

$$x_{n+1} = x_n - \theta_1. \tag{32}$$

 ISBN: 978-84-608-6082-2

By imposing the first two tangency conditions, we have

$$\theta_1 = \theta_5 f(x_n), \;\; \theta_4 = \frac{1-\theta_5 f'(x_n)}{f(x_n)}. \tag{33}$$

From the last three tangency conditions, we obtain

$$\begin{aligned}
&f(y_n)\big[f'(x_n)\left(f'(x_n)\left(2\theta_5 f'(x_n)-1\right)+\theta_3 f(x_n)^2\right)-\theta_2 f(x_n)^3\big] = f'(x_n)^2 f(x_n)\left(\theta_5 f'(x_n)-1\right),\\
&f(z_n)\left[\frac{(1-\theta_5 f'(x_n))(z_n-x_n)}{f(x_n)}+\theta_2(z_n-x_n)^3+\theta_3(x_n-z_n)^2+\theta_5\right] = \theta_5 f(x_n)+z_n-x_n,\\
&f(t_n)\left[\frac{(1-\theta_5 f'(x_n))(t_n-x_n)}{f(x_n)}+\theta_2(t_n-x_n)^3+\theta_3(t_n-x_n)^2+\theta_5\right] = \theta_5 f(x_n)+t_n-x_n.
\end{aligned} \tag{34}$$

By eliminating θ_2 and θ_3 from the above equations, we get

$$\theta_5 = \frac{a_n b_n\big(u_1 f(x_n)^2 f(y_n)+u_2 f'(x_n) f(t_n) f(z_n)\big)}{v_1 f(x_n)^3+v_2 f'(x_n) f(t_n) f(z_n)}, \tag{35}$$

where

$$\begin{aligned}
u_1 &= f(t_n)\big(b_n^2 f'(x_n)+b_n f(x_n)-c_n f(z_n)\big)+a_n\big(f(x_n)-a_n f'(x_n)\big)f(z_n),\\
u_2 &= a_n b_n c_n f'(x_n)\big(f(y_n)-f(x_n)\big)+c_n f(y_n) f(x_n)(a_n-b_n),\\
v_1 &= f(y_n)\big[b_n f(t_n)\left(b_n^2 f'(x_n)+b_n f(x_n)-c_n f(z_n)\right)+\big(a_n^3 f'(x_n)+c_n a_n f(t_n)-a_n^2 f(x_n)\big)f(z_n)\big],\\
v_2 &= a_n^2 b_n^2 c_n f'(x_n)^2\big(2f(y_n)-f(x_n)\big)+a_n b_n c_n(2a_n-c_n) f'(x_n) f(y_n) f(x_n)+c_n\big(a_n b_n-a_n c_n-b_n^2\big) f(y_n) f(x_n)^2,\\
a_n &= x_n-z_n,\; b_n = t_n-x_n,\; c_n = t_n-z_n.
\end{aligned}$$

Now, by using the equations (28), (32), (33) and (35), we obtain

$$\begin{aligned}
y_n &= x_n-\frac{f(x_n)}{f'(x_n)},\\
z_n &= y_n-\frac{f(x_n)+\beta f(y_n)}{f(x_n)+(\beta-2)f(y_n)}\frac{f(y_n)}{f'(x_n)},\\
t_n &= z_n-\frac{f(z_n)}{f'(x_n)}G(u,\,v),\\
x_{n+1} &= x_n-\theta_5 f(x_n).
\end{aligned} \tag{36}$$

where θ_5 is defined by (35). The following Theorem 3.1 demonstrates that the optimal sixteenth-order of convergence is achieved.

Theorem 3.1 *Under the assumptions of Theorem 2.2, the iterative scheme defined by* (36) *has an optimal sixteenth-order convergence and satisfies the following error equation*

$$e_{n+1} = -\frac{c_2^3((2\beta+1)c_2^2-c_3)^2}{2}\Big[c_2^4\left(4\beta^3-32\beta^2+44\beta+2\beta G_{12}+G_{12}+(2\beta+1)^2G_{20}-82\right) + c_3^2(G_{20}-2)-2c_4c_2-c_3c_2^2(-4\beta+G_{12}+4\beta G_{20}+2G_{20}-30)\Big](c_2^4-3c_3c_2^2+2c_4c_2 + c_3^2-c_5)e_n^{16}+O(e_n^{17}). \tag{37}$$

Proof Let us expand the function $f(x_n)$ and its first order derivative $f'(x_n)$ around $x=\alpha$, by the Taylor's series expansion, we have

$$f(x_n) = f'(\alpha)\left[\sum_{k=1}^{16} c_k e_n^k + O(e_n^{17})\right], \tag{38}$$

and

$$f'(x_n) = f'(\alpha)\left[\sum_{k=1}^{16} k c_k e_n^{k-1} + O(e_n^{17})\right], \tag{39}$$

respectively.
By using the equations (38) and (39), we get

$$e_{n,y} = y_n - \alpha = c_2e_n^2 - 2(c_2^2-c_3)e_n^3 + \sum_{j=1}^{13} P_k e_n^{k+3} + O(e_n^{17}), \tag{40}$$

where $P_k = P_k(\beta,\ c_2,\ c_3,\ \ldots,\ c_{16})$.
With the help Taylor series expansion, we obtain the following expansion of $f(y_n)$ about a point $x=\alpha$

$$f(y_n) = f'(\alpha)\left[\sum_{k=1}^{8} c_k e_{n,y}^k + O(e_n^{17})\right], \tag{41}$$

By using the equations (38), (39) and (41), in the second-step, we obtain

$$e_{n,z} = z_n - \alpha = \left((2\beta+1)c_2^3 - c_2c_3\right)e_n^4 + \sum_{j=1}^{12} Q_k e_n^{k+4} + O(e_n^{17}), \tag{42}$$

where $P_k = P_k(\beta,\ c_2,\ c_3,\ \ldots,\ c_{16})$.
Again, expand the Taylor series expansion of function $f(z_n)$ about a point $z=\alpha$, we obtain

$$f(z_n) = f'(\alpha)\left[\sum_{k=1}^{4} c_k e_{n,z}^k + O(e_n^{17})\right], \tag{43}$$

 ISBN: 978-84-608-6082-2

By inserting the equations (13), (14), (38)–(43) in the third sub step, we have

$$e_{t_n} = t_n - \alpha = -\frac{c_2((2\beta+1)c_2^2 - c_3)}{2}\Big[c_2^4\left(4\beta^3 - 32\beta^2 + 44\beta + 2\beta G_{12} + G_{12} + (2\beta+1)^2 G_{20} - 82\right) + c_3^2(G_{20} - 2) - 2c_4c_2 - c_3c_2^2(-4\beta + G_{12} + 4\beta G_{20} + 2G_{20} - 30)\Big]e_n^8 + \sum_{j=1}^{8} R_k e_n^{k+8} + O(e_n^{17}), \tag{44}$$

where $R_k = R_k(\beta,\ c_2,\ c_3,\ \ldots,\ c_{16})$.

Again, with the help of Taylor series, we have

$$f(t_n) = f'(\alpha)\left(e_{t_n} + c_2 e_{t_n}^2 + O(e_n^{17})\right). \tag{45}$$

Using equations (38) – (45), in the last sub step of the proposed scheme (37) and further simplifying the equations, we get

$$e_{n+1} = -\frac{c_2^3((2\beta+1)c_2^2 - c_3)^2}{2}\Big[c_2^4\left(4\beta^3 - 32\beta^2 + 44\beta + 2\beta G_{12} + G_{12} + (2\beta+1)^2 G_{20} - 82\right) + c_3^2(G_{20} - 2) - 2c_4c_2 - c_3c_2^2(-4\beta + G_{12} + 4\beta G_{20} + 2G_{20} - 30)\Big](c_2^4 - 3c_3c_2^2 + 2c_4c_2 + c_3^2 - c_5)e_n^{16} + O(e_n^{17}). \tag{46}$$

This reveals that the proposed scheme (37) reaches an optimal sixteenth-order convergence. This is the complete proof of the theorem. □

4 Numerical experiments

This section is fully devoted to check the effectiveness and validity of our theoretical results which we have proposed in the earlier Sections. For this purpose, most of the times some researchers who want to claim that their methods are superior than other existing methods available in the literature. They take some well-known or standard or self-made examples and manipulate the initial approximations to claim that their methods are superior than other methods. To halt this practice, we consider five numerical examples; first one is considered from Heydari et al. [1]; second one is chosen from Khatri [2]; third one is taken from Bi et al. [3]; fourth one is chosen from Guem and Kim [4] and last one is considered from Thukral [5], with same initial guesses which are mentioned in their papers. The details of chosen examples or test functions are available in Table 1. Moreover, the considered test functions with their corresponding zeros and initial guesses are also displayed in the same table.

First of all, we employ the new contributed eighth-order families namely, (24) (for $\beta = 0,\ -\frac{1}{2},\ -\frac{1}{3}$) and (26) (for $\beta = -\frac{1}{3}$), called by (CM_1), (CM_2), (CM_3) and (CM_4), respectively, to check the effectiveness and validity of the theoretical results. We will compare

our methods with two new families of iterative methods for solving nonlinear equations with optimal eighth-order convergence designed by Heydari et al. [1], out of which we consider one of their best method (which is claimed by them not by us) namely, method (14) (for $\lambda = 30,\ \theta = 6,\ a = 8$), denoted by (HHL). In addition, we also compare them with three-step optimal iterative methods with eighth-order convergence presented by Khatri [2] and Bi et al. [3], out which we choose methods namely, method (2.1) (for $\alpha = 56,\ \beta = -1,\ \mu = 0$) and method (36) (for $\alpha = 1$), called by (KM) and (BRW), respectively. Moreover, we will also compare our methods with a multi-parameter family of three-step eighth-order iterative methods proposed by Guem and Kim [4], out of which we consider method (YK1), called by (GK). Finally, we also compared our methods with a new eighth-order iterative method (17) for solving nonlinear equations developed by Thukral [5], denoted by (TM).

In the context of sixteenth-order methods for solving nonlinear equations, we employ the same weight functions which we have considered in the eighth-order schemes CM_i, $i = 1, 2, 3, 4$, to obtain the corresponding sixteenth-order iterative methods given by (36). We have called the new iterative scheme by $\widehat{CM_1}$, $\widehat{CM_2}$, $\widehat{CM_3}$ and $\widehat{CM_4}$, respectively. Now, we will compare our sixteenth-order methods with a non optimal sixteenth-order method (4) presented by Li et al. [6], called by $(LMMW)$. In addition, we will also consider optimal sixteenth-order methods namely, method (19) (for $a = 1$) and method (9), from the methods proposed by Sharma et al. [7] and Ullah et al. [8], denoted by (SGG) and (UFA), respectively. Moreover, we will also compare them with a family of multipoint methods for non-linear equations designed by Neta [9], out of which we consider method (12), denoted by (NM). Finally, we will also compare them with a biparametric family of optimally convergent sixteenth-order multipoint methods proposed by Geum and Kim [10], out of the proposed methods we shall choose the expression (1.7), called by $\widehat{GK}$.

Further, we have displayed the errors in the iterations $|x_n - \alpha|$, computational order of convergence [16], in the Table 2, 3. Further, where the exact root is not available then we restore the approximated root up to minimum 500 significant digits. All computations have been done in Mathematica (Version 9) with multiple precision arithmetic to minimize the round-off errors. Let us remark that, in all tables, $a\ e(\pm b)$ denotes $a \times 10^{(\pm b)}$.

5 Conclusions

In this paper, we contributed further to the development of the theory of iteration processes and presented some new novel eighth and sixteenth-order families of King's method for solving nonlinear equations which is based on the weight function and rational functional approximation approaches. Analysis of convergence demonstrate that the order of convergence of the proposed families are eight and sixteen. Further, we also prove that the proposed families are optimal in the sense of the classical Kung-Traub conjecture. The computational efficiency index is defined as $E = p^{1/\theta}$, where p is the order of convergence

 ISBN: 978-84-608-6082-2

and θ is the number of functional evaluations per iteration. Thus, the efficiency indices of the proposed families are $E = \sqrt[4]{8} \approx 1.682$ and $E = \sqrt[5]{16} \approx 1.741$ which are better than the classical Newton's method $E \approx 1.414$. Moreover, the beauty of the proposed families is that we can easily obtain several new optimal methods of order eight and sixteen by considering different types of weight functions.

Finally, on accounts of the results obtained, it can be concluded that our proposed methods are highly efficient and perform better than the existing methods of same order even though if we consider the same nonlinear problems with same initial guesses as they consider in their own papers (for the details please see the Table 1).

References

[1] M. Heydari, S.M. Hosseini, G.B. Loghmani, *On two new families of iterative methods for solving nonlinear equations with optimal order*, Appl. Anal. Dis. Math. **5** (2011) 93–109.

[2] S.K. Khatri, *Optimal eighth order iterative methods*, Math. Comput. Sci. **5** (2011) 237–243.

[3] W. Bi, H. Ren, Q. Wu, *Three-step iterative methods with eighth-order convergence for solving nonlinear equations*, J. Comput. Appl. Math. **225** (2009) 105–112.

[4] Y.H. Geum, Y.I. Kim, *A multi-parameter family of three-step eighth-order iterative methods locating a simple root*, Appl. Math. Comput. **215** (2010) 3375–3382.

[5] R. Thukral, *A new eighth-order iterative method for solving nonlinear equations*, Appl. Math. Comput. **217** (2010) 222–229.

[6] X. Li , C. Mu, J. Ma, C. Wang, *Sixteenth-order method for nonlinear equations*, Appl. Math. Comput. **215** (2010) 3754–3758.

[7] J. R. Sharma, R.K. Guha, Puneet Gupta, *Improved King's methods with optimal order of convergence based on rational approximations*, Appl. Math. Lett. **26** (2013) 473–480.

[8] M.Z. Ullah, A.S. Al-Fhaid, F. Ahmad, *Four-point optimal sixteenth-order iterative method for solving nonlinear equations*, J. Appl. Math. **2013** (2013), Article ID 850365, 5 pages.

[9] B. Neta, *On a family of multipoint methods for non-linear equations*, Int. J. Comput. Math. **9** (1981) 353–361.

[10] Y.H. Geum, Y.I. Kim, *A biparametric family of optimally convergent sixteenth-order multipoint methods with their fourth-step weighting function as a sum of a rational and a generic two-variable function*, J. Comput. Appl. Math. **235** (2011) 3178–3188.

[11] S.C. Chapra, R.P. Canale, *Numerical methods for engineers with programing and software applications*, Tata McGraw-Hill publishing company limited, New Delhi, 2000.

[12] R. Behl, S.S. Motsa, *Geometric Construction of Eighth-Order Optimal Families of Ostrowski's Method*, T. Sci. W. J. **2015** (2015) article ID 614612, 11 pages.

[13] A. Cordero, J.L. Hueso, E. Martínez, J.R. Torregrosa, *New modifications of Potra-Pták's method with optimal fourth and eighth order of convergence*, J. Comput. Appl. Math. **234** (2010) 2969–2976.

[14] A. Cordero, J.R. Torregrosa, M. P. Vassileva, *Three-step iterative methods with optimal eighth-order convergence*, J. Comput. Appl. Math. **235** (2011) 3189–3194.

[15] E. Y.H. Geum, Y.I. Kim, *A family of optimal sixteenth-order multipoint methods with a linear fraction plus a trivariate polynomial as the fourth-step weighting function*, Comput. Math. Appli. **61** (2011) 3278–3287.

[16] M. Grau-Sánchez, M. Noguera, J.M. Gutiérrez, *On some computational orders of convergence*, Appl. Math. Lett. **23** (2010) 472–478.

[17] R.F. King, *A family of fourth order methods for nonlinear equations*, SIAM J. Numer. Anal. **10** (1973) 876–879.

[18] H.T. Kung, J.F. Traub, *Optimal order of one-point and multipoint iteration*, J. ACM **21** (1974) 643–651.

[19] L. Liu, X. Wang, *Eighth-order methods with high efficiency index for solving nonlinear equations*, J. Comput. Appl. Math. **215** (2010) 3449–3454.

[20] M. S. Petković, B. Neta, L. D. Petković, and J. Džunić, *Multipoint methods for solving nonlinear equations*, Academic Press, 2012.

[21] J.R. Sharma, R. Sharma, *A new family of modified Ostrowski's methods with accelerated eighth order convergence*, Numer. Algor. **54** (2010) 445–458.

[22] F. Soleymani, S. Karimi Vanani, M. Khan, M. Sharifi, *Some modifications of King's family with optimal eighth-order of convergence*, Math. Comput. Model. **55** (2012) 1373–1380.

[23] F. Soleymani, M. Sharifi, B.S. Mousavi, *An improvement of Ostrowski's and King's techniques with optimal convergence order eight*, Opt. The. Appl. **153(1)** (2012) 225–236.

[24] R. Thukral, M.S. Petkovíc, *A family of three point methods of optimal order for solving nonlinear equations*, J. Comput. Appl. Math. **233** (2010) 2278–2284.

[25] J.F. Traub, *Iterative methods for the solution of equations*, Prentice-Hall, Englewood Cliffs, 1964.

[26] X. Wang, L. Liu, *Modified Ostrowski's method with eighth-order convergence and high efficiency index*, Appl. Math. Lett. **23** (2010) 549–554.

Table 1: Test problems

$f(x)$	$zeros(\alpha)$	x_0
$f_1(x) = (x-2)\left(x^{10}+x+1\right)e^{-x-1}$; [1]	2	2.1
$f_2(x) = x^4 + \sin\left(\frac{\pi}{x^2}\right) - 5$; [2]	$\sqrt{2}$	1.2
$f_3(x) = x^5 + x^4 + 4x^2 - 15$; [3]	1.34742809896830498150671 5...	1.6
$f_4(x) = \frac{\log\left(x^2+2x+2\right)}{x^2+1} + \left(x^2+1\right)\cos\left(\frac{\pi x}{2}\right)$; [4]	-1	-0.81
$f_5(x) = xe^{x^2} - \sin^2(x) + 3\cos(x) + 5$; [5]	$-1.20764782713091892700941 7...$	-1

Table 2: Comparison of $|x_n - \alpha|$ for the functions $f_i(x),\ i = 1,\ 2, \ldots, 5$ among listed methods

f_i	$\lvert x_n-\alpha\rvert\backslash\rho$	HHL	KM	BRW	GK	TM	CM_1	CM_2	CM_3	CM_4
	$\lvert x_1-\alpha\rvert$	$9.9e(-6)$	$9.3e(-4)$	$1.8e(-5)$	$3.1e(-5)$	$2.9e(-4)$	$2.2e(-5)$	$8.0e(-6)$	$2.0e(-6)$	$5.1e(-8)$
f_1	$\lvert x_2-\alpha\rvert$	$3.8e(-36)$	$5.0e(-19)$	$3.1e(-34)$	$5.8e(-34)$	$2.7e(-23)$	$2.8e(-34)$	$1.3e(-37)$	$3.2e(-43)$	$1.3e(-55)$
	$\lvert x_3-\alpha\rvert$	$2.0e(-279)$	$3.3e(-141)$	$2.4e(-264)$	$8.1e(-264)$	$1.4e(-175)$	$2.2e(-265)$	$8.2e(-292)$	$1.8e(-337)$	$3.0e(-436)$
	ρ	8.000	8.003	8.000	8.000	8.000	8.000	8.000	8.000	8.000
	$\lvert x_1-\alpha\rvert$	$3.1e(-6)$	$1.4e(-5)$	$3.3e(-6)$	$1.6e(-6)$	$3.8e(-6)$	$1.1e(-6)$	$5.7e(-7)$	$7.2e(-8)$	$8.3e(-6)$
f_2	$\lvert x_2-\alpha\rvert$	$8.5e(-45)$	$7.3e(-38)$	$4.4e(-44)$	$1.9e(-47)$	$2.3e(-445)$	$1.1e(-48)$	$9.5e(-51)$	$6.5e(-58)$	$9.8e(-41)$
	$\lvert x_3-\alpha\rvert$	$2.6e(-353)$	$4.3e(-296)$	$1.7e(-347)$	$5.4e(-375)$	$4.1e(-350)$	$1.1e(-384)$	$5.7e(-401)$	$2.8e(-458)$	$3.9e(-320)$
	ρ	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000
	$\lvert x_1-\alpha\rvert$	$3.7e(-6)$	$8.4e(-4)$	$1.5e(-5)$	$1.3e(-5)$	$1.3e(-4)$	$3.1e(-6)$	$6.8e(-6)$	$8.4e(-7)$	$1.2e(-6)$
f_3	$\lvert x_2-\alpha\rvert$	$1.1e(-43)$	$2.7e(-23)$	$1.1e(-38)$	$9.2e(-41)$	$3.2e(-30)$	$2.6e(-46)$	$4.9e(-42)$	$1.0e(-49)$	$6.1e(-48)$
	$\lvert x_3-\alpha\rvert$	$5.8e(-344)$	$2.9e(-279)$	$6.6e(-304)$	$7.4e(-322)$	$5.2e(-235)$	$7.0e(-367)$	$3.6e(-331)$	$5.4e(-393)$	$2.5e(-378)$
	ρ	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000
	$\lvert x_1-\alpha\rvert$	$1.5e(-6)$	$4.6e(-4)$	$3.3e(-6)$	$5.2e(-7)$	$2.4e(-5)$	$4.0e(-7)$	$1.0e(-6)$	$5.8e(-7)$	$2.4e(-6)$
f_4	$\lvert x_2-\alpha\rvert$	$3.0e(-47)$	$2.6e(-26)$	$6.0e(-45)$	$3.8e(-52)$	$.1.4e(-36)$	$1.6e(-53)$	$2.4e(-49)$	$1.1e(-52)$	$9.3e(-48)$
	$\lvert x_3-\alpha\rvert$	$7.0e(-373)$	$2.2e(-204)$	$7.7e(-355)$	$3.4e(-413)$	$1.5e(-286)$	$1.4e(-424)$	$2.3e(-390)$	$1.3e(-418)$	$5.6e(-379)$
	ρ	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000	8.000
	$\lvert x_1-\alpha\rvert$	$3.5e(-4)$	$4.2e(-1)$	$2.2e(-4)$	$2.2e(-6)$	$2.8e(-3)$	$4.2e(-6)$	$1.3e(-5)$	$3.1e(-5)$	$5.4e(-4)$
f_5	$\lvert x_2-\alpha\rvert$	$2.6e(-27)$	$5.4e(-2)$	$4.1e(-28)$	$5.8e(-45)$	$1.0e(-18)$	$1.3e(-43)$	$9.7e(-39)$	$8.7e(-36)$	$5.1e(-25)$
	$\lvert x_3-\alpha\rvert$	$2.6e(-212)$	$2.6e(-7)$	$4.8e(-218)$	$1.2e(-353)$	$3.6e(-142)$	$1.1e(-343)$	$9.6e(-304)$	$3.5e(-280)$	$3.1e(-193)$
	ρ	8.000	5.940	8.000	8.000	8.000	8.000	8.000	8.000	8.000

Table 3: Comparison of $|x_n - \alpha|$ for the functions $f_i(x),\ i = 1,\ 2, \ldots, 6$ among listed methods

f_i	$\lvert x_n-\alpha\rvert\backslash\rho$	$LMMW$	SGG	UFA	NM	$\widehat{GK}$	$\widehat{CM_1}$	$\widehat{CM_2}$	$\widehat{CM_3}$	$\widehat{CM_4}$
	$\lvert x_1-\alpha\rvert$	$2.3e(-12)$	$2.9e(-10)$	$9.0e(-9)$	$7.0e(-8)$	$5.8e(-8)$	$1.7e(-10)$	$2.0e(-11)$	$2.9e(-12)$	$7.4e(-14)$
f_1	$\lvert x_2-\alpha\rvert$	$6.5e(-180)$	$3.5e(-145)$	$6.6e(-119)$	$3.7e(-104)$	$4.2e(-106)$	$1.3e(-149)$	$1.2e(-164)$	$1.6e(-179)$	$1.3e(-204)$
	$\lvert x_3-\alpha\rvert$	$1.4e(-2860)$	$6.5e(-2304)$	$4.5e(-1881)$	$1.8e(-1644)$	$2.5e(-1676)$	$1.0e(-2375)$	$5.8e(-2616)$	$9.0e(-2856)$	$1.1e(-3256)$
	ρ	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
	$\lvert x_1-\alpha\rvert$	$1.6e(-12)$	$7.4e(-13)$	$5.6e(-10)$	$8.7e(-12)$	$8.6e(-12)$	$1.4e(-12)$	$8.5e(-13)$	$1.0e(-13)$	$1.2e(-11)$
f_2	$\lvert x_2-\alpha\rvert$	$1.0e(-189)$	$1.5e(-194)$	$1.7e(-146)$	$8.0e(-177)$	$4.9e(-176)$	$3.3e(-191)$	$7.6e(-194)$	$1.0e(-208)$	$4.7e(-175)$
	$\lvert x_3-\alpha\rvert$	$4.7e(-3025)$	$1.1e(-3101)$	$7.6e(-2331)$	$2.5e(-2817)$	$5.2e(-2804)$	$2.9e(-3049)$	$1.4e(-3090)$	$1.0e(-3328)$	$2.6e(-2789)$
	ρ	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
	$\lvert x_1-\alpha\rvert$	$1.7e(-12)$	$1.3e(-12)$	$2.9e(-9)$	$2.1e(-9)$	$2.5e(-9)$	$7.1e(-13)$	$1.3e(-12)$	$2.8e(-14)$	$4.1e(-14)$
f_3	$\lvert x_2-\alpha\rvert$	$7.0e(-190)$	$1.5e(-192)$	$5.9e(-135)$	$1.8e(-137)$	$2.8e(-136)$	$4.7e(-198)$	$2.2e(-192)$	$1.3e(-219)$	$1.4e(-216)$
	$\lvert x_3-\alpha\rvert$	$4.9e(-3028)$	$3.2e(-3071)$	$5.7e(-2146)$	$2.2e(-2186)$	$2.0e(-2167)$	$5.6e(-3161)$	$1.6e(-3068)$	$2.9e(-3505)$	$3.5e(-3456)$
	ρ	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
	$\lvert x_1-\alpha\rvert$	$6.0e(-14)$	$3.5e(-12)$	$2.9e(-9)$	$2.7e(-11)$	$2.2e(-7)$	$1.0e(-13)$	$4.5e(-13)$	$1.3e(-13)$	$5.2e(-13)$
f_4	$\lvert x_2-\alpha\rvert$	$1.1e(-215)$	$1.0e(-184)$	$3.4e(-128)$	$4.0e(-169)$	$1.4e(-107)$	$6.4e(-211)$	$4.6e(-200)$	$1.4e(-211)$	$1.0e(-201)$
	$\lvert x_3-\alpha\rvert$	$1.1e(-3443)$	$1.9e(-2945)$	$3.9e(-2039)$	$2.0e(-2694)$	$1.2e(-1710)$	$3.2e(-3366)$	$6.6e(-3192)$	$9.1e(-3379)$	$3.3e(-3221)$
	ρ	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00
	$\lvert x_1-\alpha\rvert$	$5.9e(-9)$	$2.9e(-10)$	$5.0e(-5)$	$1.1e(-7)$	$1.2e(-2)$	$1.4e(-12)$	$6.5e(-11)$	$1.0e(-10)$	$1.8e(-9)$
f_5	$\lvert x_2-\alpha\rvert$	$5.5e(-130)$	$2.6e(-152)$	$6.0e(-64)$	$1.7e(-108)$	$1.2e(-2)$	$3.3e(-191)$	$1.2e(-162)$	$1.1e(-159)$	$5.8e(-139)$
	$\lvert x_3-\alpha\rvert$	$1.7e(-2066)$	$4.1e(-2425)$	$4.8e(-1006)$	$1.4e(-1721)$	$1.2e(-2)$	$2.9e(-3049)$	$1.2e(-2590)$	$2.1e(-2543)$	$6.1e(-2211)$
	ρ	16.00	16.00	16.00	16.00	1.000	16.00	16.00	16.00	16.00

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Dynamical properties of traffic speed

Tomáš Martinovič[1]

[1] *IT4Innovations, VŠB - Technical University of Ostrava, Czech Republic*

emails: tomas.martinovic@vsb.cz

Abstract

The main aim of the paper is focused on analysis of dynamical properties of the road traffic time series, that are discrete time real data. For the analysis we use well established methods as 0-1 test for chaos and Shannon entropy. 0-1 chaos is used to determine whether the underlying dynamics of the time series is chaotic or not. Shannon entropy quantify the complexity of the system and can be used to compare complexity of different parts of the time series. As a main result strong chaotic behaviour of the real world traffic time series will be shown.

Key words: Dynamical system, traffic time series, 0-1 test for chaos, Shannon entropy, chaos
MSC 2000: 65P20, 37M10

1 Introduction

It is important to study traffic dynamics for better understanding of the underlying system, its reaction to certain events and possibility of a prediction. These information may be used for optimal traffic control systems, improving future planning of traffic infrastructure, handling existing issues in traffic management or in traffic routing problems.

In past there were research aimed at investigating chaotic behaviour in the traffic time series data. A number of works, e.g. [5, 6, 8], identify the chaos in the experimental data through the computation of maximal Lyapunov exponent. This method requires initial reconstruction of original phase-space of the attractor [10]. However, without knowing the exact mathematical formula defining the system, it is very hard to detect the correct choices of parameters determining the embedding of the time series.

In this work we analyse highway traffic dynamics of the real world time series. As indicators of characteristics of the time series we used 0-1 test for chaos and Shannon entropy. These methods were used because of suitability for basic analysis of the experimental data.

2 Tests of time series dynamics

First of all an approach for detecting chaotic behaviour in the time series called *0-1 test for chaos* will be introduced. This test gives binary result indicating whether the chaos is observable or not, and as an advantage there are no preprocessings needed comparing to the computational methods based on embedding e.g. methods used for determining of Lyapunov exponents. Secondly Shannon entropy is used to compute complexity of the given time series and its sub-sequence. Shannon entropy allows us to determine development of time series based on its complexity. Higher Shannon entropy indicate atypical events during the given time period. This information helps us to identify incidents on highways, during time series analysis, that need special attention.

2.1 0-1 test for chaos

While there were well established methodologies for identifying chaos in real world time series, however those are often based on data preprocessing. On contrary, Gottwald and Melbourne [2] responded to this by proposing a 0-1 test for chaos, which works directly with the time series. Output of this test is 0 if the underlying system of the time series is non-chaotic and 1 if the system is chaotic.

Basis for this test is creating translation variables defined for $c \in (0, \pi)$

$$p_c(n) = \sum_{j=1}^{N} x(j) \cos jc \tag{1}$$

$$q_c(n) = \sum_{j=1}^{N} x(j) \sin jc, \tag{2}$$

where $x(j)$ is the j-th element of time series x, $j \in \{1, 2, \ldots, N\}$. The system dynamics can be visualed in a plot of p versus q. Bounded plot indicates regular dynamics, while plot similar to Brownian motion indicates chaos. To quantify this problem mean square displacement have to be computed

$$M_c(n) = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} |p_c(j+n) - p_c(j)|^2 + |q_c(j+n) - q_c(j)|^2. \tag{3}$$

The mean square displacement is bounded for the regular dynamics, and has growing trend for chaotic dynamics. Since experimental data are finite, the n must be much smaller than N for the limit of convergence. It is enough to put $n = N/10$, see e.g. [3].

The oscillating part of M_c is affecting results of the 0-1 test for chaos. Therefore it is necessary to adjust it by removing the main oscillatory part of M_c. For that purpose D_c is defined as adjusted mean square displacement

$$D_c(n) = M_c - \left[\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} x_j\right]^2 \frac{1-\cos nc}{1-\cos c}. \tag{4}$$

Now, the growth rate of D_c is defined by

$$K_c = \lim_{n\to\infty} \frac{\log D_c(n)}{\log n}. \tag{5}$$

The correlation method was used to compute this limit, where D_c trend was correlated to the steadily increasing trend. This means, that vectors $t = (1, 2, \ldots, n)$ and $d = (D_c(1), D_c(2), \ldots, D_c(n))$ were created. Finally

$$K_c = \mathrm{corr}(t, d) = \frac{\mathrm{cov}(t, d)}{\sqrt{\mathrm{var}(t)\mathrm{var}(d)}} \in [-1, 1]. \tag{6}$$

This definition gives us results of 0 if the M_c is bounded and therefore the dynamics of the system are regular. Conversely the results equals 1 if the M_c is steadily increasing, implying chaotic dynamics.

It was pointed out, see e.g. [4], that computations have to be carried out for at least 100 randomly selected values of parameter $c \in (0, \pi)$, because if c is corresponding with the frequency of the time series it can provide false results. The final value of the 0-1 chaos is given by the median of K_c, since the median is not affected by the outlier values.

2.2 Shannon entropy

Shannon studied information contained in the message and quantified the expected amount of information in the message through information entropy [9] so called Shannon entropy [7].

Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be the finite set of the experiment outcomes. Variable p_i is probability of obtaining outcome a_i, $0 \le p_i \le 1$, $\sum_{i=1}^{n} p_i = 1$. Then the information entropy of $\mathcal{A}$ is

$$H(\mathcal{A}) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{7}$$

here the convention $0 \log_2 0 = 0$ is accepted.

If a partition $\mathcal{A}$ consists of k subsets, then $H(\mathcal{A})$ is bounded from above. Moreover, $H(\mathcal{A}) \le \log_2 k$, see e.g. [1]. This means that maximum of $H(\mathcal{A})$ increases as the number of subsets increases. However, if probability of outcomes created by this division is 0, $H(\mathcal{A})$ will not increase. To calculate the information entropy from the time series we created

$$k = \frac{\max(x) - \min(x)}{\min(\mathrm{diff}(x))}$$

subsets for the time series x, where $\text{diff}(x) = |x_i - x_j|$ for all $i, j \in \{1, 2, \ldots, N\}$, and $i \neq j$. The length of the interval of each subset is equal to $\min(\text{diff}(x))$. This ensures that the further division of subset would not increase the information entropy. Although Shannon entropy does not provide general information about predictability of the system, we can use it for comparison of behaviour of the same system under different conditions.

3 Results

The highway traffic data provided by Highways England [1] under the OGL license [2], have been used as experimental data. Our data sets contain average speed information on given road segment. It begins on 1st August 2013 and spans to 31st December 2014, with 15 minutes time steps. This is 49728 observations in total. Graphical display of the time series is in Figure 1.

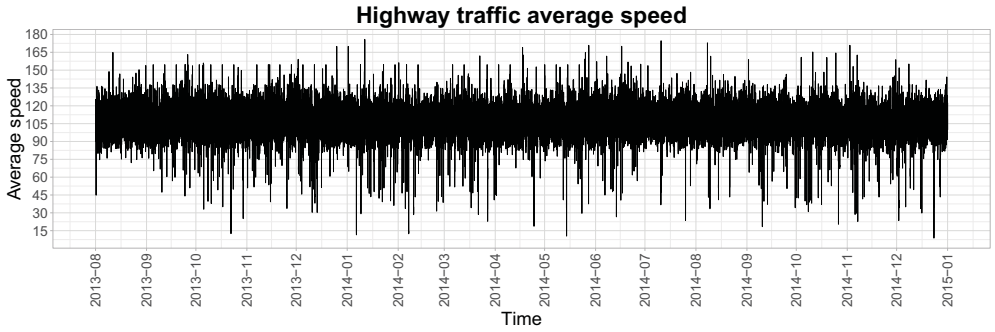


Figure 1: Traffic speed time series

Different subsets of the time series were tested to check the consistency of results. Firstly we used the whole time series for the analysis and afterwards we divided it into the months and ran the tests for each month separately. We received consistent results, where 0-1 test for chaos always returned values on the interval (0.996, 0.999), confirming chaotic behaviour of the time series. Shannon entropy shown us that the level of complexity differs during the year. In Table 1 we can see that in August 2013 the Shannon entropy was the smallest, meaning the traffic situation was most stable. The complexity of the traffic increased in March 2014 and continued until July 2014. Another important information is, that we got similar results of Shannon entropy for the same months in year 2013 and year 2014. This means the traffic dynamics are similar each year.

[1] https://data.gov.uk/dataset/dft-eng-srn-routes-journey-times
[2] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Test	2013-08	2013-09	2013-12	2014-03	2014-06	2014-09	2014-12
0-1 chaos	0.998	0.997	0.997	0.998	0.998	0.997	0.998
Shannon	9.787	9.849	9.920	10.015	10.032	9.938	9.924

Table 1: Traffic dynamics test results

4 Conclusion

In this article we analysed real world traffic speed time series. We used the 0-1 test for chaos to show that traffic speed dynamics on highways are chaotic. In addition, the Shannon entropy index illustrated different behaviour of the dynamics during the year, with increased chaotic behaviour in the second quarter of the year. In addition to that we found that the dynamics of highway traffic tends to repeat each year.

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project IT4Innovations excellence in science - LQ1602.

References

[1] G. H. Choe, *Computational Ergodic Theory*, Springer-Verlag, Berlin, 2005.

[2] G. A. Gottwald, I. Melbourne, *A new test for chaos in deterministic systems*, Proc. R. Soc. London A **460** (2004) 603–611.

[3] G. A. Gottwald, I. Melbourne, *On the implementation of the 0-1 test for chaos*, SIAM J. Appl. Dyn. Syst. **8** (2009) 129–145.

[4] G. A. Gottwald, I. Melbourne, *Testing for chaos in deterministic systems with noise*, Physica D: Nonlinear Phenomena **212** (2005) 100–110.

[5] J. Hu, Y. Wang, Z. Zhang, D. Li, *Analysis on traffic flow data and extraction of nonlinear characteristic quantities*, Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference **13** (2010) 712–717.

[6] M. Hui, L. Bai, Y. Li, Q. Wu, *Highway Traffix Flow Nonlinear Character Analysis and Prediction*, Mathematical Problems in Engineering **2015** (2015) .

[7] A. János, Z. Daróczy, *On measures of information and their characterizations* New York (1975)

[8] S. Li, G. He, *Chaos Characteristics of Expressway Traffic Flow*, J. Highway Transp. Res. Dev.(English Ed.) **2** (2007) 66–69.

[9] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal **27** (1948) 379–423.

[10] F. Takens, *Detecting strange attractors in turbulence*, Lecture Notes in Mathematics **98** (1981) 366–381.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Benchmarking of third-order reduced density matrices approximations using the harmonium atom.

Eduard Matito[1,2,3], Mauricio Rodríguez-Mayorga[2,4], Eloy Ramos-Cordoba[1,2,5] and Ferran Feixas[4]

[1] *Kimika Fakultatea, Euskal Herriko Unibertsitatea (UPV/EHU)*

[2] *Donostia International Physics Center (DIPC),*

[3] *IKERBASQUE, Basque Foundation for Science*

[4] *Institut de Química Computacional i Catàlisi (IQCC), Univ. Girona*

[5] *Kenneth S. Pitzer Center for Theoretical Chemistry, UC Berkeley*

email: ematito@gmail.com

Abstract

The computation of third-order reduced density matrices (3-RDM) are needed for the calculation of electronic energies from the contracted Schrödinger equation and the antiHermitian counterpart. In practice, approximations to these matrices are used, to avoid the large cost associated to its calculation. To our knowledge, there has not been an exhaustive study of the correlation effects in 3-RDM approximations. In this work we have used the three-electron harmonium atom —which provides a formidable playground to study electron correlation— to analyze the Valdemoro, Nakatsuji and Mazziotti approximations to the 3-RDM, and an in-house approximation for the diagonal of the 3-RDM. Our results suggest that Valdemoro's approximation is the most suited for moderate correlation effects but it also puts forward the overall poor performance of the existing approximations to deal with highly-correlated systems.

Key words: third-order reduced density matrix, harmonium atom, N-representability

1 Introduction

For a system of fermions subject to one and two-particle forces the exact energy can be completely expressed in terms of the second-order reduced density matrix (2-RDM). Many authors have attempted to calculate the ground-state energy from the 2-RDM because it is a much simpler object than the electronic wavefunction. The use of the variational method to calculate the energy of a system involves the modification of the 2-RDM subject to the N-representability conditions. Despite the progress in the quest for the complete set of N-representability conditions [1], a practical solution to the problem remains to be found. Notwithstanding, the contracted Schrödinger equation (CSE) [2, 3] and the antiHermitian counterpart (ACSE) [4] have rekindled the interest in methods without wavefunctions [5]. Both CSE and ACSE energy expressions depend on the third-order reduced density (3-RDM) that is usually approximated from lower-order densities [6–8].

2 Methodology

The diagonal part of the 3-RDM (and higher-order matrices), commonly known as the three-particle density (3-PD), can be also used to calculate the so-called n-center electron sharing indices (nc-ESI) [9], through the following formula: [a]

$$\delta(A_1, A_2, ..., A_n) = \frac{(-2)^{n-1}}{(n-1)!} \int_{A_1} d_{\mathbf{1}} \int_{A_2} d_{\mathbf{2}} \cdots \int_{A_n} d_{\mathbf{n}}\, \gamma(\mathbf{1}, \mathbf{2}, ..., \mathbf{n})\,, \tag{1}$$

where

$$\gamma(\mathbf{1}, \mathbf{2}, \ldots, \mathbf{n}) = \langle (\hat{\rho}_1 - \bar{\rho}_1)(\hat{\rho}_2 - \bar{\rho}_2) \cdots (\hat{\rho}_n - \bar{\rho}_n) \rangle\,. \tag{2}$$

and $\gamma(\mathbf{1}, \mathbf{2}, \ldots, \mathbf{n})$ involves the computation of the n-density, $\rho_n(\mathbf{1}, \ldots, \mathbf{n})$,

$$\langle \hat{\rho}_1 \cdots \hat{\rho}_n \rangle_{A_1 \ldots A_n} = \int_{A_1} d_{\mathbf{1}} \cdots \int_{A_n} d_{\mathbf{n}}\, \rho_n(\mathbf{1}, \ldots, \mathbf{n}) \tag{3}$$

and the lower-order densities in a set of three-dimensional space regions. The $\delta(A_1, \ldots, A_n)$ is invariant with respect to the order of the atoms in the string and is proportional to the n-central moment of the n-variate probability distribution $\rho_n(\mathbf{1}, \ldots, \mathbf{n})$ integrated into the

[a] In the following we will indicate the coordinates of the electron using the short-hand notation $\mathbf{1} \equiv (\vec{r}_1, \sigma_1)$ and $d_{\mathbf{1}} \equiv d\vec{r}_1 d\sigma_1$ for the derivatives. A semicolon (;) will be used to separate l.h.s. coordinates from r.h.s. coordinates. The absence of the semi-colon indicates that we refer to the diagonal part of the matrix.

 ISBN: 978-84-608-6082-2

atomic basins $A_1, \ldots, A_n$: [b]

$$\delta(A_1, \ldots, A_n) = \frac{(-2)^{n-1}}{(n-1)!} \left\langle \prod_{i=1}^{n} \left(\hat{N}_{A_i} - \overline{N}_{A_i} \right) \right\rangle \tag{4}$$

where $\hat{N}_A$ is the particle operator applied to region A and $\overline{N}_A$ is the average number of electrons in A (or population of A):

$$\bar{N}(A) = \int \hat{N}_A \rho(\mathbf{1}) d_{\mathbf{1}} \equiv \int_A \rho(\mathbf{1}) d_{\mathbf{1}} \tag{5}$$

In a recent work [10] we have put forward two new approximations to the 3-PD that have been used to calculate 3c-indices in a series of molecules. Our approximations were compared against the Valdemoro [6], Nakatsuji [7] and Mazziotti [8] approximations, showing that one of our proposals was clearly superior to the others for the calculation of 3c-indices. This expression was named cube root (CR) approximation, it reduces to the exact 3-PD for non-correlated systems and is the only approximation to attain the sum rule,

$$\rho_3^{\mathrm{CR}}(\mathbf{1}, \mathbf{2}, \mathbf{3}) = \gamma^{\mathrm{CR}}(\mathbf{1}, \mathbf{2}, \mathbf{3}) - 2\rho(\mathbf{1})\rho(\mathbf{2})\rho(\mathbf{3}) + \hat{\pi}_1^3 \, \rho_2(\mathbf{1}, \mathbf{2})\rho(\mathbf{3}) \tag{6}$$

where $\hat{\pi}_1^3$ is an operator which generates the two possible subsets of indices of sizes 1 and 2 in the set $\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$, ρ_2 is the two-particle density (2-PD) and

$$\gamma^{\mathrm{CR}}(\mathbf{1}, \mathbf{2}, \mathbf{3}) = 2 \sum_{ijk} (n_i n_j n_k)^{1/3} \, \phi_i^*(\mathbf{1}) \phi_j(\mathbf{1}) \phi_k(\mathbf{2}) \phi_j^*(\mathbf{2}) \phi_i(\mathbf{3}) \phi_k^*(\mathbf{3}) \tag{7}$$

where $\phi_i(\mathbf{1})$ is a natural orbital and n_i its occupation number. The CR approximation to the 3-PD bears a close resemblance with Müller's approximation to the 2-PD [11] and provides a simple formulae to calculate the 3c-indices only in terms of natural orbitals:

$$\tilde{\delta}^{\mathrm{CR}}(A_1, A_2, A_3) = 4 \sum_{ijk} (n_i n_j n_k)^{1/3} \, S_{ij}(A_1) S_{jk}(A_2) S_{ki}(A_3) \tag{8}$$

where $S_{ij}(A_1)$ is the atomic overlap matrix (AOM) of atom A_1, which we calculate as

$$S_{ij}(A_1) = \int_{A_1} d_{\mathbf{1}} \phi_i^*(\mathbf{1}) \phi_j(\mathbf{1}) \, . \tag{9}$$

In this work we explore the role of correlation effects into the aforementioned approximations to the 3-RDM (and the approximations of the 3-PD recently suggested). To this

[b] This expectation values do not include the self-pairing of electrons (explicitly forbidden by Pauli's principle).

 ISBN: 978-84-608-6082-2

aim, we have chosen a model system, the harmonium atom (HA) [12], where the electrons are confined on a parabolic potential, $\frac{1}{2}\omega^2\mathbf{r}^2$,

$$H = \sum_i^N \left(-\frac{1}{2}\nabla^2_{\mathbf{r}_i} + \frac{1}{2}\omega^2\mathbf{r}_i^2 \right) + \sum_{i<j}^N \frac{1}{r_{ij}} \tag{10}$$

where ω is the confinement strength. This models allows an easy tuning of the amount of correlation by playing with the ω parameter. For large values of ω electrons are in a low-correlation regime, whereas the small-ω region corresponds to highly correlated systems. For the present study we have taken the quartet ($S = 3/2$) and the doublet ($S = 1/2$) of the three-electron HA for several values of the ω parameter ($\omega \in [0.1, 1000]$). Full configuration interaction calculations on these states of the 3e-HA from a previous study [13] will be used to generate the exact 3-RDM and various approximations. The orbital representation of these matrices, ${}^3D = \{{}^3D^{ijk}_{lmn}\}$,

$$\rho(\mathbf{1}', \mathbf{2}', \mathbf{3}'; \mathbf{1}, \mathbf{2}, \mathbf{3}) = \sum_{\substack{ijk \\ lmn}} {}^3D^{ijk}_{lmn} \phi^*_i(\mathbf{1}')\phi^*_j(\mathbf{2}')\phi^*_k(\mathbf{3}')\phi_l(\mathbf{1})\phi_m(\mathbf{2})\phi_n(\mathbf{3}) \tag{11}$$

will be assessed using a series of tests: (*i*) fulfilment of the sum rule,

$$\mathrm{Tr}\left[{}^3D\right] = N(N-1)(N-2)$$

(*ii*) attainment of some well-known N-representability conditions [14] (see Fig. 1), (*iii*) calculation of 3c-indices in three arbitrary regions of the three-dimensional space and (*iv*) a termwise assessment.

Our results [15] suggest that Valdemoro's approximation is the most suited for correlation effects but it also puts forward the overall poor performance of the existing approximations to deal with highly-correlated systems.

Acknowledgements

This research has been funded by Spanish MINECO Project No. CTQ2014-52525-P and the Basque Country Consolidated Group Project No. IT588-13. The author acknowledges the computational resources and technical and human support provided by SGI/IZO-SGIker UPV/EHU.

References

[1] D. A. Mazziotti, Phys. Rev. Lett. **108**, 263002 (2012).

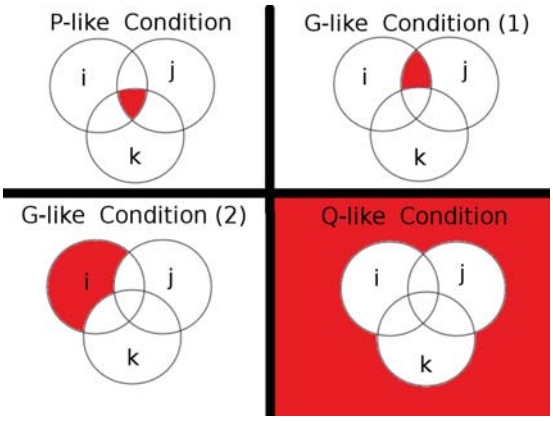


Figure 1: Sketch of some N-Representability Conditions.

[2] H. Nakatsuji, Phys. Rev. A **14**, 41 (1976).

[3] J. Cioslowski, *Many-electron densities and reduced density matrices* (Kluwer Academic, New York, 2000).

[4] D. A. Mazziotti, Phys. Rev. Lett. **97**, 143002 (2006).

[5] D. A. Mazziotti, Acc. Chem. Res. **39**, 207 (2006).

[6] F. Colmenero and C. Valdemoro, Int. J. Quant. Chem. **51**, 369 (1994).

[7] H. Nakatsuji and K. Yasuda, Phys. Rev. Lett. **76**, 1039 (1996).

[8] D. A. Mazziotti, Phys. Rev. A **60**, 4396 (1999).

[9] M. Giambiagi, M. S. de Giambiagi, and K. C. Mundim, Struct. Chem. **1**, 423 (1990).

[10] F. Feixas, M. Solà, J. M. Barroso, J. M. Ugalde, and E. Matito, J. Chem. Theory Comput. **10**, 3055 (2014).

[11] A. M. K. Müller, Phys. Lett. **105A**, 446 (1984).

[12] N. R. Kestner and O. Sinanoglu, Phys. Rev. **128**, 2687 (1962).

[13] J. Cioslowski and E. Matito, J. Chem. Theory Comput. **7**, 915 (2011).

[14] F. Weinhold and E. B. Wilson Jr, The Journal of Chemical Physics **47**, 2298 (1967).

[15] M. Rodríguez-Mayorga, E. Ramos-Cordoba, F. Feixas, and E. Matito (in preparation).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Consensus formation in a system of difference equations modeling controversial opinion dynamics with pairwise interactions

M. G. Medina-Guevara[1], J. E. Macías-Díaz [2], A. Gallegos[1] and H. Vargas-Rodríguez[1]

[1] *Departamento de Ciencias Exactas y Tecnología, Centro Universitario de los Lagos, Universidad de Guadalajara*

[2] *Departamento de Matemáticas y Física, Universidad Autónoma de Aguascalientes*

emails: MGuadalupe.Medina@lagos.udg.mx, jemacias@correo.uaa.mx, gallegos@culagos.udg.mx, hvargas@culagos.udg.mx

Abstract

In this work, we propose a discrete system to model the dynamics of opinions on controversial subjects. The social network consists of a finite number of agents with pairwise interactions at discrete times. Meanwhile, the opinion of each agent is updated following a general nonlinear law which considers parameters identified as the personal cultural baggages of each of the members. We establish conditions that guarantee the existence of global attracting points (strong consensus) and intervals (weak consensus). Moreover, we notice that these conditions are independent of the weight matrix and the number of agents in the network. One particular scenario modeling extremist and centrist postures is investigated numerically in order to confirm the validity of the analytical results.

Key words: Consensus formation, controversial opinion dynamics, stability analysis, strong and weak consensus

1 Introduction

The specialized literature has already proposed various mathematical models to investigate the dynamics of opinions in social networks, some of the most popular being the models proposed by Sznajd-Weron and Sznajd [8], the voter model in complex networks [7], the

models investigated by Deffuant *et al.* [1] and the Hegselmann–Krause system [3]. In all of these models, the opinion of each agent is a consequence of interactions between members of the social network (for example, such interactions are considered in the weights matrix of the De Groot model [2], and Hegselmann–Krause system). Moreover, the determination of conditions that give rise to consensus has been a common topic of many research articles, including those cited above.

It is important to point out that the notion of consensus in sociology has been compared in relevance to the concept of energy in physics [6], and the investigation of the existence of consensus has been carried out on discrete spaces, voter models and bounded confidence systems alike. However, when personal values and beliefs are directly involved in an opinion, the postures are often defensive and the opinions steady, and many of the classical models become inapplicable in such circumstances. In view of those limitations, the purpose of this work is to investigate the formation of consensus in a new model that describes the dynamics of opinions on controversial topics.

Opinion dynamics has been frequently modeled using fully discrete dynamical systems [4]. Indeed, realistic settings actually consider the presence of a finite number of agents as well as discrete interactions within finite subgroups of the social network at discrete instants of time, whence the use of systems of coupled difference equations is justified in the practice. In the present study, we will propose a system of difference equations to model the dynamics of opinions on controversial subjects. Our model may conveniently incorporate attractors and repellers in order to describe strong postures of the agents, together with an individual parameter proposed previously in [9]. The inclusion of that additional parameter gives the advantage of mimicking the existence of strong personal attitudes and values often motivated by hard changing cultural paradigms in each of the agents of the network. In that work, the author justified its use by social and/or environmental reasons, but we will employ it in the present manuscript to prescribe the cultural factors as well as personal values, knowledge, predisposition, etc. In summary, the parameter used in [9] will be employed here to represent strong attitudes acquired during the lifetime of each individual, and we will denominate it the *cultural baggage* or, simply, the *baggage* of the agent.

2 Preliminaries

2.1 Nomenclature

Throughout this manuscript, we will suppose that the temporal interval $[0, \infty)$ is partitioned into a (not necessarily uniform) sequence of points $0 = t_0 < t_1 \ldots < t_n < \ldots$, for $n \in \mathbb{Z}^+ \cup \{0\}$. We will assume that X is a connected subset of $\mathbb{R}$ that represents the *space of opinions*. Meanwhile, the set $K \subseteq \mathbb{R}$ will represent the collection of possible values of the cultural baggages of the agents, and it will be called the *space of baggages*. In this work, we will fix an even number N of agents conforming the population under investigation,

 ISBN: 978-84-608-6082-2

and let $\Omega_N = \{1, 2, \ldots, N\}$. At the nth temporal step, the ith agent will be completely characterized by an ordered pair $\mathbf{x}_n^i = (x_n^i, \kappa^i)$ consisting of an opinion $x_n^i \in X$ and a baggage $\kappa^i \in K$, for each $i \in \Omega_N$.

Following the approach of [2], the social network under consideration will be described by the $N \times N$ square matrix of absolute weights, c_{ij}.

For each pair of indexes $i, j \in \Omega_N$, the number c_{ij} will represent the trust that the ith agent has on the opinion of the jth agent. The weights do not vary in time, and must satisfy the condition $\sum_{j=1}^{N} c_{ij} = 1$, for each $i \in \Omega_N$. Moreover, we will assume that the inequalities $0 \leq c_{ij} < 1$ and $0 < c_{ii} \leq 1$ hold for any pair of different indexes $i, j \in \Omega_N$.

Throughout, we let ϵ be a positive number which will be called the *coefficient of affinity* or the *confidence level* [3], and it will be used to restrict the interaction between the agents of the population. We say that the opinions of the ith and the jth agents at time t_n are *affine* if $|x_n^i - x_n^j| < \epsilon$, for each $i, j \in \Omega_N$ and $n \in \mathbb{Z}^+ \cup \{0\}$. Equivalently, we say simply that $\mathbf{x}_n^i$ and $\mathbf{x}_n^j$ are affine.

Let $\Xi : X \times K \to \mathbb{R}$ used to model a process of internal reflection for the individual opinion update of the agents of the population according to their values and attitudes. For analytical reasons, we may assume that X is an open interval of $\mathbb{R}$. Under these conditions, the function Ξ will be differentiable in the first variable.

2.2 Recursive model

We propose an iterative model in order to simulate the dynamics of the opinions of the agents. At the initial time, we provide an initial stochastic matrix c_{ij} and set initial conditions for each member of the population. Following the approaches in [1], only binary interactions between members are allowed. Thus, at each iteration at most $N/2$ pairs of agents are randomly chosen to exchange opinions. Let $n \in \mathbb{Z}^+ \cup \{0\}$ and let $i, j \in \Omega_N$ be such that $i \neq j$. We consider two possible scenarios, depending on the affinity of the opinions of $\mathbf{x}_n^i$ and $\mathbf{x}_n^j$:

- **Opinions are affine.** Under these circumstances, the opinions for the $(n+1)$st temporal step will be defined through

$$\begin{cases} x_{n+1}^i &= a_{ii}\Xi(x_n^i, \kappa^i) + a_{ij}x_n^j, \\ x_{n+1}^j &= a_{ji}x_n^i + a_{jj}\Xi(x_n^j, \kappa^j), \end{cases} \quad \text{where} \quad \begin{array}{ll} a_{ii} = \frac{c_{ii}}{c_{ii}+c_{ij}}, & a_{ij} = \frac{c_{ij}}{c_{ii}+c_{ij}}, \\ a_{ji} = \frac{c_{ji}}{c_{jj}+c_{ji}}, & a_{jj} = \frac{c_{jj}}{c_{jj}+c_{ji}} \end{array} \tag{1}$$

  Clearly, a_{ii} and a_{ij} represent respectively the relative weight that $\mathbf{x}_n^i$ gives to her own opinion, and to the opinion of $\mathbf{x}_n^j$. Obviously, $a_{ii} + a_{ij} = 1$, $0 < a_{ii} \leq 1$ and $0 \leq a_{ij} < 1$. Similar remarks can be drawn on a_{jj} and a_{ji}.

- **Opinions are not affine.** In this case, the new opinions are updated considering individual baggages exclusively. In other words, we let

$$x_{n+1}^i = \Xi(x_n^i, \kappa^i), \qquad x_{n+1}^j = \Xi(x_n^j, \kappa^j). \tag{2}$$

The agents interacting through these iterative rules together with the particular parametric values and initial conditions, will be referred to as a *social network*. Throughout, we will suppose that the network is totally connected in the sense that it has no isolated components.

Acknowledgements

This work was partially supported by the National Council for Science and Technology of Mexico (CONACYT).

References

[1] G. Deffuant, F. Amblard, G. Weisbuch, T. Faure, *How can extremism prevail? A study based on the relative agreement interaction model*, Journal of Artificial Societies and Social Simulation **5** (2002), p.1.

[2] M. H. De Groot, *Reaching a Consensus* J. Amer. Statist. Assoc. **69** (1974), pp. 118 121.

[3] R. Hegselmann, and U. Krause *Opinion dynamics and bounded confidence: models, analysis and simulation* Journal of Artificial Societies and Social Simulation **5** (2002), p.1.

[4] U. Krause *Arithmethic-geometric discrete systems*, Journal of Difference equations and Applications **12** (2006), pp. 229–231.

[5] F.R. Marotto, *Introduction to mathematical modeling using discrete dynamical systems*, Thomson Brooks/Cole, London, 2006.

[6] T.J. Scheff, *Toward a sociological model of consensus*, American Sociological Review **32** (1967), pp. 32–46.

[7] K. Suchecki, V.M. Eguíluz, and M. San Miguel, *Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution*, Physical Review E **72** (2005), p. 036132.

[8] K. Sznajd-Weron and J. Sznajd, *Opinion evolution in closed community*, Journal of Modern Physics C **11** (2000), 1157–1165.

[9] T. Yamano, *An opinion formation dynamics with logistic map and its complexity*, WSEAS Transactions on Business and Economics **1** (2004), pp.229-234.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A superconvergent partial differential equation approach to price variance swaps

Dilloo Mehzabeen Jumanah[1] **and Tangman Désiré Yannick**[1]

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

emails: mehzabeen.dilloo1@umail.uom.ac.mu, y.tangman@uom.ac.mu

Abstract

In this paper, we present a simple and fast finite difference algorithm to price variance swaps under the Black–Scholes model and complex models, like the Merton's jump-diffusion model and the two-dimensional Heston's stochastic volatility model. The associated terminal conditions due to the two most popular definitions of realised variance that exist can be translated to second-order polynomials for which the third and higher order derivatives are zero. We could therefore implement only a central second-order discretisation coupled with an exponential time integration (ETI) scheme to price variance swaps with high accuracy. For both definitions of the realised variance under the Merton's model, we represent the integral part of the partial integro-differential equation (PIDE) as a solution to an ordinary differential equation in order to efficiently apply the ETI scheme. We also fill in the gap in literature for this model by providing the analytical solution for variance swap when the realised variance is based on actual returns.

Key words: variance swaps, finite difference, exponential time integration, Merton's jump-diffusion model, Heston's stochastic volatility model
MSC 2000: 35G61, 91B25, 65M06

1 Introduction

The most popular volatility derivative that exists in financial markets is the variance swap. Replicated by a portfolio of vanilla options [1, 2], the contract has a zero upfront premium with payoff

$$V_T = L\left(\sigma_R^2 - K\right),$$

 ISBN: 978-84-608-6082-2

at maturity T, where L is the notional amount per volatility squared. At initiation of the contract,

$$V_0 = \mathbb{E}\left[\exp\left(-\int_0^T r dt\right)\left(\sigma_R^2 - K\right)\right],$$

and consequently, for constant interest rate r, K is the strike that satisfies

$$K = \mathbb{E}\left[\sigma_R^2\right], \tag{1}$$

where σ_R^2 is the realised variance that is defined in the contract. The realised variance is proportional to the sum of the squared returns over the N sampling dates and usually takes one of the most popular definitions given as

$$\sigma_{Rs}^2 = \frac{AF}{N}\sum_{i=1}^{N}\left(\underbrace{\frac{S_{t_i} - S_{t_{i-1}}}{S_{t_{i-1}}}}_{\text{Actual return}}\right)^2 100^2, \quad \sigma_{Rx}^2 = \frac{AF}{N}\sum_{i=1}^{N}\left(\underbrace{\ln\left[\frac{S_{t_i}}{S_{t_{i-1}}}\right]}_{\text{Log return}}\right)^2 100^2,$$

where AF is the annualisation factor, N is the number of observations made at regular interval of $\Delta t = 1/AF$ during the lifetime T of the variance swap and S_{t_i} is the asset price at time $t_i = i\Delta t$ for $i = 0, \cdots, N$.

The growing body of literature is repleted with closed form solutions for variance swaps. On one side for σ_{Rx}^2, Broadie and Jain [3] solved stochastic differential equations (SDEs) to reap solutions for variance swaps under various popular models like the Black–Scholes, the Merton's jump-diffusion and the Heston's stochastic volatility models. Moreover, in [4], another approach to use Fourier transform was proposed to price variance swaps under the Heston's model. However, the solutions under the Heston's model are complicated to implement and this induced the authors in [5] to propose a much simpler analytical solution. On the other side, Little and Pant [6] derived analytical solutions when using σ_{Rs}^2 under the Black–Scholes model followed by the authors in [7] who applied Fourier transform to a set of PDE problems to derive the solution for variance swaps under the Heston's model. The derivations of the solution proposed for σ_{Rs}^2 was also improved in [8] by replacing the lengthy Fourier transform approach with a simple tower property to evaluate the expectation in (1). However, closed-form solutions for variance swap under the Merton's jump-diffusion model with realised variance σ_{Rs}^2 have not yet been proposed and indeed, this forms part of the work presented in this paper. From the above literature review, we can deduce that there exists no single analytical method that can price variance swaps with the two definitions of realised variance under all pricing models.

Though Monte Carlo simulations provide a universal approach [7, 9] this method, although simple, is plagued by its low convergence rate and its high computational cost. More recently, Fourier methods have been proposed in [10] and [11]. The authors in [10]

 ISBN: 978-84-608-6082-2

considered only one-dimensional models and the extension of the proposed technique to two-dimensional models is not an easy and a straightforward task. The authors in [11] derived an approximation to the characteristic function of the realised variance under a generalised pure diffusion two-dimensional model. However, it would be more difficult to extend his method to price variance swaps with σ_{Rs}^2 under jump-diffusion models. Consequently, we aim at using a PDE approach which provides a universal technique due to the existence of pricing PDEs for all models. Little and Pant [6] have set the ground work PDE approach for pricing variance swaps with σ_{Rs}^2 but they used a Crank–Nicolson scheme with a central finite difference scheme in space to solve log-transformed PDEs which resulted in large errors. Although their scheme is second-order convergent, it requires a high number of nodes to reach accuracy even up to two decimal places. In [12], the authors have built on the approach of Little and Pant [6] and propose a high-order compact scheme to solve the system of PDEs. Although this technique improved the convergence rate, yet it will still need many computational nodes to reach its highest level of accuracy, especially under the Heston's model where conditions on the mesh size have to be observed. In [13], the authors devised a new PDE approach that can deal with non-linearity in the payoffs functions but the technique consists of increasing the dimension of the problem in space. Since the payoffs of variance swaps are linear functions in the realised variance, we consider the PDE approach architected by Little and Pant [6].

In this work, we demonstrate how to further improve the PDE approach of Little and Pant [6] by developing an efficient finite difference algorithm to price variance swaps based on the actual or logarithmic definition of the realised variance under the Black–Scholes, Merton's and Heston's models such that very high accuracy is reached using few computational grid nodes only. We also derived a closed-form solution for variance swap based on σ_{Rs}^2 under the Merton's model and also show that our result from the PDE approach converges to that solution very quickly.

The paper has the following structure. In Section 2, we describe the different models used and give the respective PDEs satisfied by a contingent claim on the asset price. In Section 3, we derive the analytical solution for variance swaps with σ_{Rs}^2 under the Merton's model to fill the gap in litterature. Then, in Section 4, we briefly outline the PDE approach used to price variance swaps and in Section 5, we give the approximations to our partial derivatives together with a description of the ETI scheme. Section 6 presents our numerical results and we conclude in Section 7.

2 Models

In this section, we give the pricing PDEs under the different models we use to price variance swaps. We also state the PDEs when the transformation $x = \ln S$ is applied since they will be useful for pricing variance swaps with realised variance σ_{Rx}^2.

2.1 The Black–Scholes model

The model assumes the following asset price dynamics

$$\frac{dS_t}{S_t} = rdt + \sigma dW_t,$$

[14] where r is the interest rate, σ is the volatility and W_t is the Wiener process responsible for the stochastic behaviour of the asset price S_t. Then, standard pricing arguments and Ito's lemma on a claim $P(t;\, S)$ which depends on the asset price S and time t leads to the following pricing PDE

$$\frac{\partial P}{\partial t} + \mathcal{L}_s = 0, \tag{2}$$

where

$$\mathcal{L}_s = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP.$$

Further, we consider the transformation $x = \ln S$ so that the claim $P(t;\, x)$ satisfies PDE (2) with $\mathcal{L}_s$ replaced by

$$\mathcal{L}_x = \frac{1}{2}\sigma^2 \frac{\partial^2 P}{\partial x^2} + \left(r - \frac{1}{2}\sigma^2\right)\frac{\partial P}{\partial x} - rP.$$

2.2 The Merton's jump-diffusion model

This model enriches the Black–Scholes model by allowing the possibility of jumps which occur according to a Poisson process such that both the jump and the Poisson process are uncorrelated to the Wiener process. Therefore, we have

$$\frac{dS_t}{S_t} = (r - \lambda_J\zeta)dt + \sigma dW_t + (Y-1)dN_t, \tag{3}$$

[15] where the Poisson process N_t has a rate λ_J. Here, $Y-1$ is an impulse function making S jump to SY and Y is taken from the log-normal distribution with probability density function

$$\tilde{f}(\tilde{y}) = \frac{1}{\sqrt{2\pi}\sigma_J\tilde{y}} \exp\left(\frac{-(\ln\tilde{y} - \mu_J)^2}{2\sigma_J^2}\right), \quad \tilde{y} > 0,$$

such that the expectation of $Y-1$ given by $\zeta = \exp\left(\mu_J + \sigma_J^2/2\right) - 1$.

Ito's formula for jump-diffusion processes [16] and hedging arguments lead us to the PIDE

$$\frac{\partial P}{\partial t} + \underbrace{\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} + (r - \lambda_J\zeta)S\frac{\partial P}{\partial S} - (r+\lambda_J)P}_{\mathcal{L}_s} + \lambda_J\mathcal{I}_s = 0,$$

 ISBN: 978-84-608-6082-2

where

$$\mathcal{I}_s = \int_0^\infty P(t;\, S\tilde{y})\tilde{f}(\tilde{y})d\tilde{y}. \tag{4}$$

Applying transformation $x = \ln S$ and accordingly, $y = \ln \tilde{y}$, we have

$$\frac{\partial P}{\partial t} + \underbrace{\frac{1}{2}\sigma^2\frac{\partial^2 P}{\partial x^2} + \left(r - \frac{1}{2}\sigma^2 - \lambda_J\zeta\right)\frac{\partial P}{\partial x} - (r+\lambda_J)P}_{\mathcal{L}_x} + \lambda_J\mathcal{I}_x = 0,$$

where

$$\mathcal{I}_x = \int_{\Re} P(t;\, z)f(z-x)dz, \tag{5}$$

$z = x + y$ and $f(y) = \exp(y)\tilde{f}(\exp(y))$.

2.3 The Heston's stochastic volatility model

This model accounts for the stochastic behaviour of volatility on the market and assumes the following coupled stochastic differential equations (SDEs)

$$\frac{dS_t}{S_t} = rdt + \sqrt{v_t}dW_t, \quad dv_t = \kappa(\theta - v_t)dt + \sigma_v\sqrt{v_t}dW_t^v,$$

[17] where v_t is the variance of the asset, κ is the mean reversion rate to the long run mean θ, and σ_v denotes the volatility of the variance. With ρ being the correlation coefficient between the two Wiener processes W_t and W_t^v, our claim $P(t;\, S,\, v)$ satisfies PDE (2) with the operator

$$\mathcal{L}_s = \frac{1}{2}S^2v\frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP + \frac{1}{2}\sigma_v^2 v\frac{\partial^2 P}{\partial v^2} + \kappa(\theta - v)\frac{\partial P}{\partial v} + \rho\sigma_v Sv\frac{\partial^2 P}{\partial S\partial v},$$

and the transformation $x = \ln S$ gives the operator

$$\mathcal{L}_x = \frac{1}{2}v\frac{\partial^2 P}{\partial x^2} + \left(r - \frac{1}{2}v\right)\frac{\partial P}{\partial x} - rP + \frac{1}{2}\sigma_v^2 v\frac{\partial^2 P}{\partial v^2} + \kappa(\theta - v)\frac{\partial P}{\partial v} + \rho\sigma_v v\frac{\partial^2 P}{\partial x\partial v},$$

where we can now see that the coefficients are independent of x but remain dependent on v.

3 Analytical solutions under the Merton's model

In this section, we propose analytical solutions for pricing variance swaps when considering discretely and continuously monitored actual returns for the realised variance under the Merton's jump-diffusion model.

To start with, suppose that the asset price jumps from S to SY_k for N_T times in the time interval $[0,\, T]$. Then the continuously monitored realised variance is given as

$$\sigma_{RC}^2 = 100^2 \frac{AF}{N} \left[\int_0^T \sigma^2 dt + \sum_{k=1}^{N_T} \left(\frac{SY_k - S}{S} \right)^2 \right],$$

so that the fair value of the strike K_C is expressed as

$$K_C = \mathbb{E}\left[\sigma_{RC}^2\right] = 100^2 \frac{AF}{N} \left(\sigma^2 T + \mathbb{E}\left[\sum_{k=1}^{N_T} (Y_k - 1)^2 \right] \right).$$

Since Y follows a lognormal distribution with mean μ_J and variance σ_J^2, and N_T is a Poisson process with mean $\lambda_J T$, the expectation on the right hand side of the above equation can be obtained from

$$\mathbb{E}\left[\sum_{k=1}^{N_T} (Y_k - 1)^2 \right] = \mathbb{E}[N_T] \mathbb{E}\left[Y_k^2 - 2Y_k + 1 \right],$$

and further, as $AF/N = 1/T$, we have

$$K_C = 100^2 \left(\sigma^2 + \lambda_J \left[\exp\left(2\mu_J + 2\sigma_J^2\right) - 2\exp\left(\mu_J + \frac{1}{2}\sigma_J^2\right) + 1 \right] \right). \tag{6}$$

We proceed by considering the corresponding discretely sampled variance swap where the fair value of the strike K_D can be obtained from

$$K_D = \mathbb{E}\left[\sigma_{Rs}^2\right] = \mathbb{E}\left[100^2 \frac{AF}{N} \sum_{i=1}^{N} \left(\frac{S_{t_i} - S_{t_{i-1}}}{S_{t_{i-1}}} \right)^2 \right],$$

and since each element in the above summation is independent of one another, we can express K_D as

$$K_D = 100^2 \frac{AF}{N} N \mathbb{E}\left[\left(\frac{S_{\tilde{t}} - S_{\tilde{t}-\Delta t}}{S_{\tilde{t}-\Delta t}} \right)^2 \right] = 100^2 AF \left[\mathbb{E}\left(\frac{S_{\tilde{t}}}{S_{\tilde{t}-\Delta t}} \right)^2 - 2\mathbb{E}\left(\frac{S_{\tilde{t}}}{S_{\tilde{t}-\Delta t}} \right) + 1 \right]. \tag{7}$$

In order to obtain the expectations in (7), we solve the SDE (3) from time $\tilde{t} - \Delta t$ to $\tilde{t}$ so that

$$\frac{S_{\tilde{t}}}{S_{\tilde{t}-\Delta t}} = \exp\left(\left(r - \frac{1}{2}\sigma^2 - \lambda_J \zeta \right) \Delta t + \sigma \left(W_{\tilde{t}} - W_{\tilde{t}-\Delta t} \right) + \sum_{k=1}^{N_{\Delta t}} \ln Y_k \right). \tag{8}$$

The careful eye will notice that the solution in (8) has the Markov property, where $S_{\tilde{t}}/S_{\tilde{t}-\Delta t}$ depends only on the Δt step. Hence, we are able to calculate the moments of $S_{\tilde{t}}/S_{\tilde{t}-\Delta t}$ with the first one being

$$\mathbb{E}\left[\frac{S_{\tilde{t}}}{S_{\tilde{t}-\Delta t}}\right] = \sum_{i\geq 0}\mathbb{P}(N_{\Delta t}=i)\times\mathbb{E}\left[\exp\left(\left(r-\frac{1}{2}\sigma^2-\lambda_J\zeta\right)\Delta t+\sigma\left(W_{\tilde{t}}-W_{\tilde{t}-\Delta t}\right)+\sum_{k=1}^{i}\ln Y_k\right)\right]. \tag{9}$$

Now, since the exponent in (9) follows a normal distribution with mean and variance given as

$$\mu_i = \left(r-\frac{1}{2}\sigma^2-\lambda_J\zeta\right)\Delta t + i\mu_J, \quad \sigma_i^2 = \sigma^2\Delta t + i\sigma_J^2,$$

respectively, and $N_{\Delta t}$ being a Poisson process with mean $\lambda_J\Delta t$, we have

$$\mathbb{E}\left[\frac{S_{\tilde{t}}}{S_{\tilde{t}-\Delta t}}\right] = \sum_{i\geq 0}\frac{\exp\left(-\lambda_J\Delta t\right)\left(\lambda_J\Delta t\right)^i}{i!}\exp\left(\mu_i+\frac{1}{2}\sigma_i^2\right). \tag{10}$$

Similarly, for the second moment of $S_{\tilde{t}}/S_{\tilde{t}-\Delta t}$, we have

$$\mathbb{E}\left[\left(\frac{S_{\tilde{t}}}{S_{\tilde{t}-\Delta t}}\right)^2\right] = \sum_{i\geq 0}\frac{\exp\left(-\lambda_J\Delta t\right)\left(\lambda_J\Delta t\right)^i}{i!}\exp\left(2\mu_i+2\sigma_i^2\right). \tag{11}$$

We notice that the infinite series in (10) and (11) are highly convergent and evaluating only a few terms in the summation before replacing them in (7), we can obtain highly accurate values for the fair value of the strike K_D.

In Appendix A, we show that our solution for the discretely sampled variance swaps tends to that for the continuously sampled one in (6) as $AF\to\infty$.

4 The partial differential equation framework

Having obtained the PDEs under the different models, we move on to describe the PDE approach in [6] to price a variance swap. For illustration, we take into consideration σ_{Rs}^2 under the Black–Scholes model noting that similar procedures are to be observed for σ_{Rx}^2 and other pricing models.

The pricing problem resides in finding the fair value of the strike K given as

$$K = \mathbb{E}\left[\sigma_{Rs}^2\right] = \frac{AF}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{S_{t_i}-S_{t_{i-1}}}{S_{t_{i-1}}}\right)^2\right]100^2.$$

Therefore, one can see that for a fixed time step Δt and $\tilde{t}=t_i=i\Delta t$, we need to find N expectations of the form

$$\mathbb{E}\left[\left(\frac{S_{\tilde{t}}-S_{\tilde{t}-\Delta t}}{S_{\tilde{t}-\Delta t}}\right)^2\right]. \tag{12}$$

 ISBN: 978-84-608-6082-2

Since at t_1, the expectation to be found involves only one unknown $S_{\Delta t}$, and at t_i for $i = 2$ to N, the latter consists of two unknowns $S_{\tilde{t}-\Delta t}$ and $S_{\tilde{t}}$, the pricing problem is divided into two cases.

4.1 For case $\tilde{t} = i\Delta t$, where $i = 2$ to N

First, we consider the case when $i = 2$ to N, where we have the two unknowns $S_{\tilde{t}-\Delta t}$ and $S_{\tilde{t}}$. We follow the steps in [6] where the authors introduce a new variable I_t which jumps at $t = \tilde{t} - \Delta t$ with the following definition

$$I_t = \int_0^t \delta\left(\tilde{t} - \Delta t - \tau\right) S_\tau \, d\tau = \begin{cases} 0, & 0 \le t < \tilde{t} - \Delta t, \\ S_{\tilde{t}-\Delta t}, & t \ge \tilde{t} - \Delta t. \end{cases}$$

Now, our contingent claim P also depends on I and consequently satisfies the following PDE

$$\frac{\partial P}{\partial t} + \mathcal{L}_s + \delta(t - \tilde{t} - \Delta t)\frac{\partial P}{\partial I} = 0,$$

for $t \in [0, \tilde{t}]$. Then, to reduce the dependence on I in the above PDE, we exploit the dirac delta function, $\delta(t - \tilde{t} - \Delta t)$, which is zero away from $\tilde{t} - \Delta t$. Consequently, for $t \in [0, \tilde{t}]\backslash(\tilde{t} - \Delta t)$ our pricing PDE is reduced to

$$\frac{\partial P}{\partial t} + \mathcal{L}_s = 0. \tag{13}$$

Therefore, solving PDE (13) subject to the terminal condition[1]

$$P(\tilde{t};\, S,\, I) = \left(\frac{S}{I} - 1\right)^2, \tag{14}$$

and jump condition

$$P(\tilde{t} - \Delta t^-;\, S^-,\, I^-) = P(\tilde{t} - \Delta t^+;\, S^+,\, I^+),$$

which, by definition of I_t, simplifies to

$$P(\tilde{t} - \Delta t^-;\, S^-,\, 0) = P(\tilde{t} - \Delta t^+;\, S^+,\, S^+), \tag{15}$$

we can obtain $P(0;\, S_0)$ so that the expectation in (12) can be calculated using Feynman–Kac theorem from

$$\mathbb{E}\left[\left(\frac{S_{\tilde{t}} - S_{\tilde{t}-\Delta t}}{S_{\tilde{t}-\Delta t}}\right)^2\right] = \exp\left(r\tilde{t}\right) P(0;\, S_0).$$

[1] $I_t = S_{\tilde{t}-\Delta t}$ for $t > \tilde{t} - \Delta t$

 ISBN: 978-84-608-6082-2

Here, we remark that a change of variable $\bar{P}(0;\, S_0) = \exp\left(r\tilde{t}\right) P(0;\, S_0)$ implies that the reaction term in $\mathcal{L}_s$ vanishes, that is, the term rP disappears.

The approach of Little and Pant [6] is not straightforward to implement due to the jump condition (15) that has to be satisfied for each grid node in the computational domain

$$\Omega_{\Delta S} = \left\{ S_j \in \Re^+ : S_j = S_{\min} + (j-1)\Delta S \right\},$$

for $j = 1, \cdots, n+1$ and $\Delta S = (S_{\max} - S_{\min})/n$.

From the above discussion, we deduce that we need to go through the following two stages to evaluate (12):

1. Solve for $P(\tilde{t} - \Delta t;\, S_j,\, S_j)$ for each j by solving PDE (13) subject to terminal condition (14) with $I = S_j$ when $t \in (\tilde{t} - \Delta t,\, \tilde{t}]$. Here, we optimize the approach to find $P(\tilde{t} - \Delta t;\, S_j,\, S_j)$ by extending the domain with ΔS such that for each j, $S_j \in [S_{\text{minext}},\, S_{\text{maxext}}]$. Therefore, our terminal condition indexed by $I = S_j$ from $\Omega_{\Delta S}$ can be represented as a matrix of size $(n_1 + 1) \times (n + 1)$, where $n_1 + 1$ is the number of nodes in the extended domain. Hence, the PDE (13) is solved once instead of $n + 1$ times through the use of loops as in [6]. The piece of Matlab® code in Appendix B illustrates how we obtain the values for $P(\tilde{t} - \Delta t;\, S_j,\, S_j)$ for each j when pricing variance swaps based on σ^2_{Rs}.

2. Then use $P(\tilde{t} - \Delta t;\, S_j,\, S_j)$ to implement the jump condition in (15) to obtain $P(\tilde{t} - \Delta t;\, S_j,\, 0)$ which represents the terminal condition for the problem defined on the time domain $[0,\, \tilde{t} - \Delta t]$. Therefore, we can solve for $P(0;\, S_0)$ using PDE (13) when $t \in [0,\, \tilde{t} - \Delta t]$ with terminal condition $P(\tilde{t} - \Delta t;\, S_j,\, 0)$.

4.2 For case $\tilde{t} = \Delta t$ when $i = 1$

In this case $S_{\tilde{t} - \Delta t} = S_0$ is known and we only need to solve the PDE (13) subject to the terminal condition

$$P(\Delta t;\, S) = \left(\frac{S - S_0}{S_0} \right)^2, \tag{16}$$

in order to calculate the desired expectation as

$$\mathbb{E}\left[\left(\frac{S_{\Delta t} - S_0}{S_0} \right)^2 \right] = \exp\left(r\Delta t\right) P(0;\, S_0) = \bar{P}(0;\, S_0).$$

We point out that for the boundary conditions of our PDE problems, we follow the authors in both [6] and [13] to set $\partial P/\partial S = 0$ as well as $\partial^2 P/\partial S^2 = 0$ at both boundaries.

The key steps in our algorithm can be summarised as follows and we provide the Matlab® code used to price variance swaps under the Black–Scholes model in Appendix B.

 ISBN: 978-84-608-6082-2

if $i = 1$ **then**
Solve PDE for $t \in [0,\, \Delta t]$ with terminal condition (16) and store $\bar{P}_1(0;\, S_0)$.
else
Solve PDE for $t \in [(i-1)\Delta t,\, i\Delta t]$ with terminal condition (14) indexed by $I = S_j$, represented in matrix form to obtain $P((i-1)\Delta t;\, S_j,\, S_j)$ for all j.
(step is same for all $i > 1$ since PDE is solved on a Δt step with the same terminal condition.)
for $i = 2$ to N **do**
Solve PDE on for $t \in [0,\, (i-1)\Delta t]$ with terminal condition $P((i-1)\Delta t;\, S,\, S)$ and store $\bar{P}_i(0;\, S_0)$.
end for
end if
Calculate the fair value of a variance swap as $\frac{100^2 AF}{N} \sum_{i=1}^{N} \bar{P}_i(0;\, S_0)$.

5 Finite difference approximations

In this section, we give the central second-order spatial approximations and show how the errors to these approximations are eliminated. We also illustrate how to efficiently discretise the integral parts in (4) and (5) in the S- and x-directions respectively. Finally, we propose the use of the exact in time scheme which has no temporal discretisation error to price variance swaps.

5.1 Spatial discretisations

We consider the computational grid $\Omega_{\Delta S}$ and use the following central second-order approximations

$$\begin{aligned}
\left.\frac{\partial P}{\partial S}\right|_{S=S_j} &\approx \quad \tfrac{-P_{j-1}+P_{j+1}}{2\Delta S} - \tfrac{\Delta S^2}{6}\tfrac{\partial^3 P}{\partial S^3}(S_j) + \mathcal{O}(\text{higher order terms}),\\
\left.\frac{\partial^2 P}{\partial S^2}\right|_{S=S_j} &\approx \quad \tfrac{P_{j-1}-2P_j+P_{j+1}}{\Delta S^2} - \tfrac{\Delta S^2}{12}\tfrac{\partial^4 P}{\partial S^4}(S_j) + \mathcal{O}(\text{higher order terms}).
\end{aligned}$$

As such for $j = 1, \cdots, n+1$, we have the following two tridiagonal matrices

$$D_1^s = \frac{1}{2\Delta S}\text{tridiag}\left[-1 \quad 0 \quad 1\right], \quad D_2^s = \frac{1}{\Delta S^2}\text{tridiag}\left[1 \quad -2 \quad 1\right], \tag{17}$$

which respectively approximates the first and second order derivatives of P.

To demonstrate how we can obtain very high accuracy from these simple discretisations, we analyse the terminal conditions (14) of our pricing problem. They can be represented as shown in Table 1 where $x = \ln S$ and accordingly $u = \ln I$.

 ISBN: 978-84-608-6082-2

Realised variance	S-direction	x-direction
σ^2_{Rs}	$P_s(\tilde{t};\, S) = \left(\frac{S}{I} - 1\right)^2$	$P_s(\tilde{t};\, x) = (\exp(x-u) - 1)^2$
σ^2_{Rx}	$P_x(\tilde{t};\, S) = \left[\ln\left(\frac{S}{I}\right)\right]^2$	$P_x(\tilde{t};\, x) = (x-u)^2$

Table 1: Terminal conditions in the S- and x- directions for the different realised variances.

It can be easily seen that

$$\frac{\partial^n P_s(S)}{dS^n} = \frac{\partial^n P_x(x)}{dx^n} = 0 \quad \text{for} \quad n \geq 3,$$

but

$$\frac{\partial^n P_s(x)}{dx^n} \neq 0 \quad \text{and} \quad \frac{\partial^n P_x(S)}{dS^n} \neq 0 \quad \text{for} \quad n \geq 3.$$

Hence, pricing variance swaps with σ^2_{Rs} in the S-direction and with σ^2_{Rx} in the x-direction, would give exact values for D_1 and D_2 in (17) as the truncation errors in the central approximations are expressions in the third and higher order derivatives, which all completely vanish. This simple observation helps us to attain the highest level of accuracy when pricing variance swaps under the various pricing models considered in this work.

5.2 Discretisation of the integral part

We now consider the integral part in (5) which is the solution at $t = \sigma_J^2/2$ to the following PDE problem [18]

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{2\mu_J}{\sigma_J^2}\frac{\partial U}{\partial x},\ -\infty \leq x \leq \infty,\ 0 \leq t \leq \frac{\sigma_J^2}{2}, \tag{18}$$

with $P(t;\, z)$ as the initial condition.

Using the central differences in (17) to approximate the partial derivatives in (18), we are left with an ordinary differential equation which can easily be solved to give

$$\mathcal{I}_x = \exp\left(\frac{\sigma_J^2}{2}\left[D_2^x + \frac{2\mu_J}{\sigma_J^2}D_1^x\right]\right)P(t;\, x), \tag{19}$$

at $t = \sigma_J^2/2$ given that $z \equiv x$ on the grid structure.

So far in literature, the transformation $x = \ln S$ has been usually applied to solve the integral in (4). However, to reach high accuracy in space when using σ^2_{Rs}, we need to discretise the integral in the S-direction itself. Therefore, to obtain an efficient approximation for $\mathcal{I}_s$ in (4), we apply the reverse transformation $S = \exp(x)$ to (18) so that

$$\mathcal{I}_s = \exp\left(\frac{\sigma_J^2}{2}\left[I_{s^2}D_2^s + \left(1 + \frac{2\mu_J}{\sigma_J^2}\right)I_s D_1^s\right]\right)P(t;\, S), \tag{20}$$

where I_s represents a diagonal matrix containing the grid nodes in S.

5.3 Exponential time integration (ETI) scheme

Applying a time reversed transformation, $\tau = T - t$, we solve the resulting semi-discretised system

$$\frac{dP}{d\tau} = \mathbf{L}P,$$

where $\mathbf{L}$ denotes the approximation to the operators in either $\mathcal{L}_s$, $\mathcal{L}_x$, $\mathcal{L}_s + \mathcal{I}_s$ or $\mathcal{L}_x + \mathcal{I}_x$, to yield

$$P(\tau + \Delta\tau) = \exp(\mathbf{L}\Delta\tau)\, P(\tau), \tag{21}$$

for a time step $\Delta\tau$. The matrix exponential in (19), (20) and (21) can be calculated using the in-built function `expm` in Matlab®. In case large matrices are involved, Carathéodory Fejér approximations can be used to evaluate the matrix exponential more efficiently. We refer the readers to [19, 20] for more details of the implementation but as we shall see next, only a few computational nodes in space are necessary to reach a high accuracy level. Therefore, only the exponentiation of matrices with moderate sizes is required.

6 Numerical experiments

In this section, we present and analyse the results of our numerical experiments. Our proposed approach consists of pricing variance swaps based on σ^2_{Rs} in the S-direction and those based on σ^2_{Rx} in the x-direction using central second-order spatial discretisations with the ETI scheme (ETI-2). We refer our readers to [6, 7] for exact solutions when pricing with σ^2_{Rs} under the Black–Scholes and Heston's models respectively while numerical solutions for σ^2_{Rx} are compared to analytical solutions from [3] under the Black–Scholes and Merton's models, and from [5] under the Heston's model. We also validate our closed-form solution under the Merton's model for σ^2_{Rs} in (7). First, we price variance swaps under the Black–Scholes model with parameters from [6] given as $AF = 52$, $T = 1$, $r = 0.05$ and $\sigma = 0.25$.

	Little & Pant approach [6]		Our approach	
n	error	time (s)	error	time (s)
10	436.1037	0.1457	1.5740e-5	0.0538
20	99.0002	0.2113	1.9694e-7	0.0653
40	24.1584	0.4806	5.9903e-9	0.0891
80	6.0031	1.0649	6.7928e-10	0.1533
ref.	627.0607860638488			

Table 2: Error and CPU time in seconds when pricing variance swaps under the Black–Scholes model for σ^2_{Rs} using Little and Pant [6] approach and our approach.

 ISBN: 978-84-608-6082-2

In the first place, we record in Table 2 and compare the results of our approach with those of Little and Pant's where variance swaps based on σ^2_{Rs} has been priced. The smaller errors and the lower CPU timings in seconds clearly demonstrate the superiority of our approach over the one proposed by Little and Pant [6]. This confirms the efficiency behind using the ETI scheme instead of the Crank–Nicolson time stepping scheme and employing appropriate spatial variables (S or $x = \ln S$) for which the higher order derivatives of the payoff vanish.

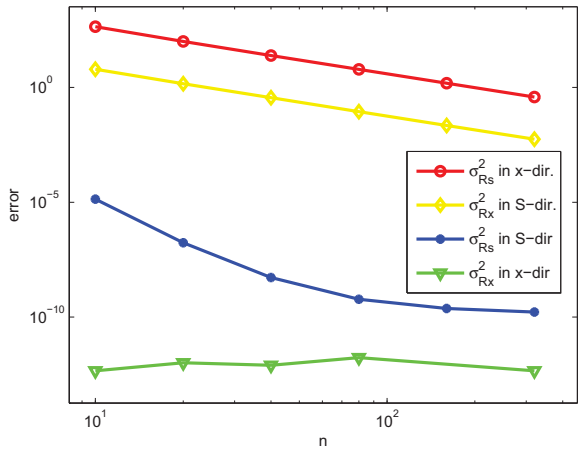


Figure 1: log(error) against n when pricing variance swaps under the Black–Scholes model for both σ^2_{Rx} and σ^2_{Rs} in the x-direction and S-direction.

Furthermore, from Figure 1, we can observe that when pricing variance swaps for both σ^2_{Rs} in the S-direction and for σ^2_{Rx} in the x-direction, superconvergence is seen as expected from the discussion in Section 5.1. Small errors of magnitude $e-10$ are obtained with about $n = 100$ computational nodes for σ^2_{Rs}. In the case of σ^2_{Rx}, approximations closer to machine precision are obtained with only $n = 10$ computational nodes. This is due to the fact that the log transformation leads to a flattening of the eigenvalues of the matrix $\mathbf{L}$ in (21) such that faster convergence is observed. This illustrates the rapidity and accuracy of our approach. Indeed, for σ^2_{Rx} in the S-direction and for σ^2_{Rs} in the x-direction, the slopes indicate that only a second-order convergence is attained and the errors are much larger than those of our approach.

	σ^2_{Rs}		σ^2_{Rx}	
n	error	time (s)	error	time (s)
10	1.2815e-8	0.0181	1.9133e-6	0.0159
20	2.6546e-10	0.0228	1.1170e-11	0.0197
40	8.6942e-11	0.0308	4.2633e-13	0.0249
ref.	166.9786595800638		176.4945892650060	

Table 3: Error and CPU time in seconds when pricing variance swaps under the Merton's model for σ^2_{Rs} in the S-domain and σ^2_{Rx} in the x-domain.

We move on to price under the Merton's jump-diffusion model. For parameters taken from [3]—$AF = 12$, $T = 1$, $\sigma = 0.1139$, $r = 0.0319$, $\lambda_J = 0.11$, $\mu_J = -0.14$ and $\sigma_J = 0.15$—we record the errors in Table 3 for both σ^2_{Rs} in the S-direction and σ^2_{Rx} in the x-direction. We can observe here that errors for σ^2_{Rx} goes up to $e-13$ and also for σ^2_{Rs}, a minimum error of $8.6942e-11$ is reached. This confirms the reliability of the proposed closed-form formula in (7) and the efficiency of our approach to discretise the integral in the S-direction as stated in Section 5.2. Moreover, Figure 2 shows that our solution for the discretely sampled variance swap in (7) converges to that for the continuously sampled one in (6) as we increase AF, the number of observations in one year.

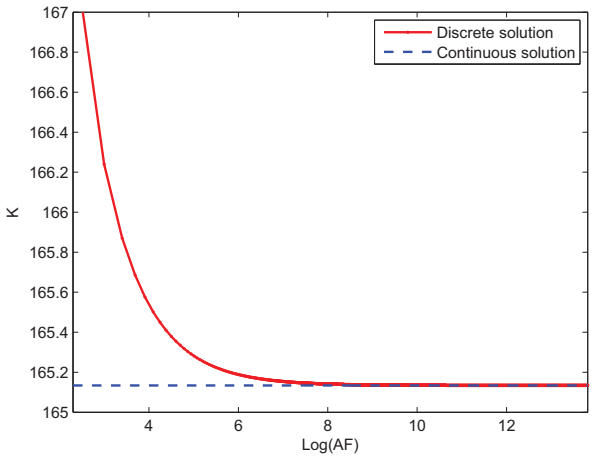


Figure 2: Convergence of the discretely sampled variance swap to the continuously sampled one as AF increases for σ^2_{Rs} under the Merton's model.

Under the Heston's model, we price using the parameters from [7] given as $AF = 52$,

$v_0 = 0.04$, $\sigma_v = 0.618$, $T = 1$, $r = 0.1$, $\kappa = 11.35$, $\rho = -0.64$ and $\theta = 0.022$. For fixed $l = 40$, the number of intervals in the v-domain, Table 4 shows that super convergence is achieved in both cases. But the error stagnates at $e-8$ for σ^2_{Rs} while a much smaller error of $e-11$ is obtained for σ^2_{Rx}.

σ^2_{Rs}			σ^2_{Rx}		
n	error	time (s)	n	error	time (s)
10	1.7609e-8	2.2166	10	4.3778e-8	2.3429
20	1.7783e-8	4.3816	20	3.3538e-11	5.4183
40	1.7780e-8	9.4997	40	3.6749e-11	13.5757
$n = 10, l$=160	6.5856-10	7.3967	$n = 20, l$=10	2.2936e-11	1.6751
ref.	237.1097368799061		ref.	238.2097413082660	

Table 4: Error and CPU time in seconds when pricing variance swaps under the Heston's model for σ^2_{Rs} in the S-domain and for σ^2_{Rx} in the x-domain. We take $l = 40$ for $v \in [0, 0.2]$.

Under the Heston's model, the solutions have another independent variable v. A close look at the analytical solution [5, 7] which consists of exponential functions in v for σ^2_{Rs} and of quadratic functions in v for σ^2_{Rx} reveals that

$$\frac{d^n Q_s}{dv^n} \neq 0, \quad \text{and} \quad \frac{d^n Q_x}{dv^n} = 0 \quad \text{for} \quad n \geq 3,$$

where $Q_s(v)$ and $Q_x(v)$ represent the analytical solutions for σ^2_{Rs} and σ^2_{Rx} respectively. This means that one would require more nodes along the v-direction for the computation with σ^2_{Rs} than with σ^2_{Rx}. This is verified in Table 4 where we use only $l = 10$ intervals in the v-direction for σ^2_{Rx} to reach the round-off plateau accuracy while we need $l = 160$ to achieve an accuracy up to $e-10$ for σ^2_{Rs}.

7 Conclusion

In this paper, we have used a simple central second-order finite difference scheme in space to price variance swaps based on actual and logarithmic returns under the Black–Scholes, the Merton's jump-diffusion and the Heston's stochastic volatility models in a single framework. By solving a set of PDEs with the spatial variable carefully chosen to match the given definition of the realised variance in the variance swap contract, we show that the truncation errors derived from the Taylor's series of the spatial approximations vanish. Hence, combining the exponential time integration which is exact in time gives the round-off plateau accuracy using a few computational grid nodes.

For the Merton's model, we propose a new efficient way of treating the integral term for the realised variance under the untransformed asset price S- spatial direction and fill the gap in the literature by giving the analytical solutions for discrete and continuous variance swaps when actual returns are considered for the realised variance. Therefore, we obtain convergent algorithms which are highly accurate within only a few seconds of computations.

A future work would be to implement the proposed approach to price variance swaps under more complex models like regime switching models under which analytical solutions are difficult to derive.

Acknowledgements

The research of Mehzabeen Jumanah Dilloo was supported by a postgraduate research scholarship from the University of Mauritius.

References

[1] K. Demeterfi, E. Derman, M. Kamal and J. Zou, *More than you ever wanted to know about volatility swaps*, Quantitative Strategies Research Notes, Goldman Sachs (1999).

[2] A. Neuberger, *Volatility trading*, London Business School working paper (1990).

[3] M. Broadie and A. Jain, *The effect of jumps and discrete sampling on volatility and variance swaps*, International Journal of Theoretical and Applied Finance. **11** (2008) 761–797.

[4] S. P. Zhu and G. H. Lian, *On the valuation of variance swaps with stochastic volatility*, Applied Mathematics and Computation. **219** (2012) 1654–1669.

[5] S. Rujivan and S. P. Zhu, *A simple closed-form formula for pricing discretely-sampled variance swaps under the Heston model*, ANZIAM Journal. **56** (2014) 1–27.

[6] T. Little and V. Pant, *A finite-difference method for the valuation of variance swaps*, Journal of Computational Finance. **5** (2001) 81–101.

[7] S. P. Zhu and G. H. Lian, *A closed-form exact solution for pricing varaince swaps with stochastic volatility*, Mathematical Finance. **21** (2011) 233–256.

[8] S. Rujivan and S. P. Zhu, *A simplified analytical approach for pricing discretely-sampled variance swaps with stochastic volatility*, Applied Mathematics Letters. **25** (2012) 1644–1650.

[9] P. Rostan, A. Rostan, A. A. E. Trach and S. Mercier, *Pricing variance and volatility swaps: a Monte Carlo simulation technique benchmarked to two closed-form solutions*, IEB International Journal of Finance. **5** (2012) 2–27.

[10] W. D. Zheng and Y. K. Kwok, *Fourier transform algorithms for pricing and hedging discretely sampled exotic variance products and volatility derivatives under additive processes*, Journal of Computational Finance. **18** (2014) 3–30.

[11] G. H. Lian, C. Chiarella and P. S. Kalev, *Volatility swaps and volatility options on discretely sampled realized variance*, Journal of Economic Dynamics and Control. **47** (2014) 239–262.

[12] M. J. Dilloo and Y. D. Tangman, *High-order compact scheme for pricing variance swaps*, Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering. **2** (2014) 466–479.

[13] H. Windcliff, P. A. Forsyth and K. R. Vetzal, *Pricing methods and hedging strategies for volatility derivatives*, Journal of Banking & Finance. **30** (2006) 409–431.

[14] F. Black and M. Scholes, *The pricing of options and other corporate liabilities*, Journal of Political Economy. **81** (1973) 637–654.

[15] R. Merton, *Theory of rational option pricing*, Bell Journal of Economics and Management Science. **4** (1973) 141–183.

[16] R. Cont and P. Tankov, *Financial modelling with jump processes*, Chapman & Hall/CRC, Boca Raton, Florida, USA (2004).

[17] S. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options*, Review of Financial Studies. **6** (1993) 327–343.

[18] P. Carr and A. Mayo, *On the numerical evaluation of option prices in jump processes*, European Journal of Finance. **13** (2007) 353–372.

[19] N. L. Trefethen, J. A. C. Weideman and T. Schmelzer, *Talbot quadratures and rational approximations*, BIT Numerical Mathematics. **46** (2006) 653–670.

[20] D. Y. Tangman, A. Peer, N. Rambeerich and M. Bhuruth, *Fast simplified approaches to asian option pricing*, Journal of Computational Finance. **4** (2011) 3–36.

 ISBN: 978-84-608-6082-2

A Limit as $AF \to \infty$ of the discretely sampled variance to the continuously sampled one under the Merton's model.

We need to prove that

$$\lim_{\Delta t \to 0} K_D = K_C,$$

since $AF = 1/\Delta t$. The above relationship holds since, firstly, it can be clearly seen that

$$\begin{aligned}\lim_{\Delta t \to 0} \quad & \left(\frac{1}{\Delta t} \underbrace{\exp\left[\left(-\lambda_J + 2\left(r - \frac{1}{2}\sigma^2 - \lambda_J \zeta \right) + 2\sigma^2 \right) \Delta t \right]}_{i=0 \text{ in } (11)} \right. \\ & \left. -2\frac{1}{\Delta t} \underbrace{\exp\left(\left(-\lambda_J + r - \frac{1}{2}\sigma^2 - \lambda_J \zeta + \frac{1}{2}\sigma^2 \right) \Delta t \right)}_{i=0 \text{ in } (10)} + \frac{1}{\Delta t} \right) \\ & = \lambda_J + \sigma^2,\end{aligned}$$

after applying Taylor series expansion to the exponential functions. Secondly, we have

$$\begin{aligned}\lim_{\Delta t \to 0} \quad & \left(\frac{1}{\Delta t} \underbrace{\lambda_J \Delta t \exp\left(-\lambda_J \Delta t + 2\left(r - \frac{1}{2}\sigma^2 - \lambda_J \zeta + \sigma^2 \right) \Delta t + 2\mu_J + 2\sigma_J^2 \right)}_{i=1 \text{ in } (11)} \right. \\ & \left. -2\frac{1}{\Delta t} \underbrace{\lambda_J \Delta t \exp\left(-\lambda_J \Delta t + \left(r - \frac{1}{2}\sigma^2 - \lambda_J \zeta + \frac{1}{2}\sigma^2 \right) \Delta t + \mu_J + \frac{1}{2}\sigma_J^2 \right)}_{i=1 \text{ in } (10)} \right) \\ & = \lambda_J \exp\left(2\mu_J + 2\sigma_J^2 \right) - 2\lambda_J \exp\left(\mu_J + \tfrac{1}{2}\sigma_J^2 \right),\end{aligned}$$

due to the cancellation of the Δt's from AF and the summations and finally,

$$\begin{aligned}\lim_{\Delta t \to 0} \quad & \left(\frac{1}{\Delta t} \sum_{i \geq 2} \frac{\exp(-\lambda_J \Delta t)(\lambda_J \Delta t)^i}{i!} \exp\left(2\mu_i + 2\sigma_i^2 \right) \right. \\ & \left. -2\frac{1}{\Delta t} \sum_{i \geq 2} \frac{\exp(-\lambda_J \Delta t)(\lambda_J \Delta t)^i}{i!} \exp\left(\mu_i + \tfrac{1}{2}\sigma_i^2 \right) \right) = 0,\end{aligned}$$

since Δt is raised to the power of $i - 1 > 0$ for each i.

 ISBN: 978-84-608-6082-2

B Matlab® code for pricing variance swaps under the Black–Scholes model with σ_{Rs}^2.

```
AF = 52;  Texp = 1; S0 = 1100; r = 0.05; sigma = 0.25;
N = AF*Texp; deltaT = 1/AF; T = [2*deltaT:deltaT:Texp];
%construction of original domain
smin = 500 ;smax = 3500; ds = [smax-smin]/n0; vecs = [smin:ds:smax]';
%construction of extended domain
m1 = ceil(400/ds); m2 = ceil(600/ds);
vecsext = [[smin-(m1*ds):ds:smin-ds]';vecs;[smax+ds:ds:smax+(m2*ds)]'];
n =length(vecsext)-1;
%construction of matrices on extended domain
e = ones(n+1,1); I = speye(n+1,n+1);
Ds = spdiags([-1*e 0*e e],-1:1,n+1,n+1)/(2*ds);
Dss = spdiags([e -2*e e],-1:1,n+1,n+1)/(ds^2);
Ds(1,:)=0; Ds(end,:)=0; Dss(1,:)=0; Dss(end,:)=0;
L = 0.5*sigma^2*spdiags(vecsext.^2,0,n+1,n+1)*Dss + r*spdiags(vecsext,0,n+1,n+1)*Ds ;
% Solution at T = deltaT on extended domain
v = ((vecsext/S0)-1).^2; v = expm(L*deltaT)*v;
expectation(1) = interp1(vecsext,v,S0,'cubic');
%Solution for T>deltaT
%Solution Stage 1
[vecss,vecii]=meshgrid(vecsext,vecs);
v = (vecss./vecii-1).^2; v = v'; v = expm(L*deltaT)*v; v = v';
H = v(find(vecss==vecii));
% Solution Stage 2
ii=[find(vecsext==vecs(1))  find(vecsext==vecs(end))];
%boundary conditions
L(ii(1),ii(1):ii(1)+1) = [0 0]; L(ii(2),ii(2)-1:ii(2)) = [0 0];
L11 = L(ii(1):ii(end),ii(1):ii(end));
for k = 1:length(T)
  v = H; v = expm(L11*(T(k)-deltaT))*v;
  expectation(k+1) = interp1(vecs,v,S0,'cubic');
end
fairvalue = 100^2* (AF/N) * sum(expectation);
```

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Sequences of sums of squares and Catalan numbers

Pedro J. Miana[1] and Natalia Romero[2]

[1] *Departamento de Matemáticas, I.U,M.A., Universidad de Zaragoza, Spain*

[2] *Departamento de Matemáticas y Computación, Universidad de La Rioja, Spain*

emails: pjmiana@unizar.es, natalia.romero@unirioja.es

Abstract

In this paper we prove new equalities involving the sequences $(a(n))_{n\geq 0}$ and $(b(n))_{n\geq 0}$ where

$$a(n):=\sum_{k=0}^{n+1}\binom{n+k}{n}^2, \qquad b(n):=\sum_{k=0}^{n+1}\frac{n+1-k}{n+1}\binom{n+k}{n}^2 \qquad n\in\mathbb{N}\cup\{0\},$$

and the well-known Catalan numbers $(C_n)_{n\geq 0}$, given by $C_n=\frac{1}{n+1}\binom{2n}{n}$, for $n\geq 0$.

In Theorems 2.1 and 2.2 we show that.

$$\begin{aligned} 2(2n+1)a(n)-na(n-1) &= (4+21n+36n^2+21n^3)C_n^2, \qquad n\geq 1;\\ 2(2n+1)b(n)-nb(n-1) &= (7n^2+8n+2)C_n^2, \qquad n\geq 1. \end{aligned}$$

As a consequence, a new proof of the nice equality $((n+1)C_n)^2=3a(n-1)-2b(n)$, for $n\geq 1$, which illustrates the intensive connection between these sequences $(a(n))_{n\geq 1}$, $(b(n))_{n\geq 1}$ and Catalan numbers, is shown.

Key words: Catalan numbers; Combinatorial identities; Binomial coefficients
MSC 2000: 05A19; 05A10; 11B65

1 Introduction

The famous Catalan numbers $(C_n)_{n\geq 0}$ counts the number of ways to triangulate a regular polygon with $n+2$ sides; or, the number of ways that $2n$ people seat around a circular table are simultaneously shaking hands with another person at the table in such a way that none of the arms cross each other, and also in tree enumeration problem, see these examples and others in [6].

There are several ways to define Catalan numbers, one of them is recursively by $C_0 = 1$ and $C_n = \sum_{i=0}^{n-1} C_i C_{n-1-i}$ for $n \geq 1$; first terms in this sequence are $1,\ 1,\ 2,\ 5,\ 14,\ 42,\ 132, \ldots$

The Catalan sequence is probably the most frequently encountered sequence and been treated deeply in many books, monographs and papers (see for example [1]-[2], [4]-[6]).

In this paper, we treat in detail the sequences $(a(n))_{n\geq 0}$ and $(b(n))_{n\geq 0}$ defined in the abstract. In Theorems 2.1 and 2.2 we show that.

$$\begin{aligned} 2(2n+1)a(n) - na(n-1) &= (4+21n+36n^2+21n^3)C_n^2, \qquad n \geq 1; \\ 2(2n+1)b(n) - nb(n-1) &= (7n^2+8n+2)C_n^2, \qquad n \geq 1. \end{aligned}$$

Lemma 2.3 shows that sequences $(a(n))_{n\geq 1}$ and $(b(n))_{n\geq 1}$ are deeply connected. Recurrence relations (2.1) and (2.2) (and polynomials in these relations) play delicate roles which allow to give an alternative proof of the beautiful equality

$$((n+1)C_n)^2 = 3a(n-1) - 2b(n), \qquad n \geq 1,$$

which was shown in [3, Theorem 3.3 (ii)]. Finally we want to point out that the WZ-theory (see details in [7]) gives computer proofs of these equalities. However analytic proofs give an extra information about the nature of these sequences which remains hidden in computer proofs.

2 Sums of squares of combinatorial numbers

We consider the sequence of integer numbers defined by

$$a(n) := \sum_{k=0}^{n} \binom{n+k}{n}^2, \qquad n \in \mathbb{N} \cup \{0\}.$$

Note that $a(0) = 1$, $a(1) = 5$, $a(2) = 46$, $a(3) = 517$, $a(4) = 6376...$. This sequence appears indexed in the On-Line Encyclopedia of Integer Sequences by N.J.A. Sloane ([5]) with the reference $A112029$. V. Kotesovec in 2012 proved the following recurrence relation

$$p_1(n)a(n) = p_2(n)a(n-1) + p_3(n)a(n-2), \qquad n \geq 2, \tag{2.1}$$

where polynomials $(p_i)_{i\in\{1,2,3\}}$ are defined by

$$\begin{aligned} p_1(n) &:= 2(2n+1)(21n-13)n^2 \\ p_2(n) &:= 1365n^4 - 1517n^3 + 240n^2 + 216n - 64, \\ p_3(n) &:= -4(n-1)(2n-1)^2(21n+8). \end{aligned}$$

Theorem 2.1. *For $n \geq 1$, the following identity holds*

$$2(2n+1)a(n) - na(n-1) = (4 + 21n + 36n^2 + 21n^3)C_n^2, \qquad n \geq 1.$$

Now we consider this second sequence of integer numbers defined by

$$b(n) := \sum_{k=0}^{n} \frac{k}{n} \binom{2n-k-1}{n-1}^2 = \sum_{k=0}^{n} \frac{n-k}{n} \binom{n-1+k}{n-1}^2, \qquad n \in \mathbb{N}.$$

Note that $b(1) = 1$, $b(2) = 3$, $b(3) = 19$, $b(4) = 163$, $b(5) = 1625...$. This sequence also appears indexed in the On-Line Encyclopedia of Integer Sequences by N.J.A. Sloane ([5]) with the reference $A183069$ and V. Kotesovec proved the following recurrence relation

$$q_1(n)b(n) = q_2(n)b(n-1) + q_3(n)b(n-2), \qquad n \geq 3, \tag{2.2}$$

where polynomials $(q_i)_{i\in\{1,2,3\}}$ are defined by

$$\begin{array}{rcl} q_1(n) & := & 2n^2(2n-1)(7n^2-20n+14) \\ q_2(n) & := & 455n^5 - 2427n^4 + 4850n^3 - 4406n^2 + 1728n - 216, \\ q_3(n) & := & -4(n-2)(2n-3)^2(7n^2-6n+1). \end{array}$$

Theorem 2.2. *For $n \geq 1$, the following identity holds*

$$2(2n+1)b(n) - nb(n-1) = (7n^2 + 8n + 2)C_n^2, \qquad n \geq 0.$$

Lemma 2.3. *For $n \geq 1$, the following two identities hold*

$$\begin{array}{rcl} \begin{vmatrix} q_1(n) & q_3(n) \\ p_1(n-1) & p_3(n-1) \end{vmatrix} & = & -8Q(n)(2n-1)(2n-3)^2(n-2); \\ \begin{vmatrix} q_1(n) & q_2(n) \\ p_1(n-1) & p_2(n-1) \end{vmatrix} & = & 16Q(n)(2n-1)(2n-3)^3, \end{array}$$

where $Q(n) := 147n^4 - 546n^3 + 666n^2 - 293n + 34$.

3 A second proof

The main aim of this section is to show an alternative of the following nice equality

$$\binom{2n}{n}^2 = \sum_{k=0}^{n} \frac{3n-2k}{n} \binom{2n-1-k}{n-1}^2, \tag{3.1}$$

 ISBN: 978-84-608-6082-2

in Theorem 3.1. As it is commented in the Introduction, the original proof is presented in [3, Theorem 3.3 (ii)] and it is a straightforward consequence of a more general equality in combinatorial numbers ([3, Theorem 3.3 (i)]). The proof which we present here allows to recognize the natural connection among the sequences $(a(n))_{n\geq 1}$, $(b(n))_{n\geq 1}$ and the Catalan numbers $(C_n)_{n\geq 1}$. Note that, once may rewrite the equality (3.1) in the following way:

$$((n+1)C_n)^2 = 3a(n-1) - 2b(n), \qquad n \geq 1.$$

Theorem 3.1. *For $n \geq 1$, the following equality holds*

$$\binom{2n}{n}^2 = \sum_{k=0}^{n} \frac{3n-2k}{n} \binom{2n-1-k}{n-1}^2.$$

Acknowledgements

Authors thank Prof. Hideyuki Ohtsuka at Bunkyo University High School in Saitama (Japan) his advice and comments to develop this alternative proof shown in the third section.

P. J. Miana has been partially supported by Project MTM2013-42105-P, DGI-FEDER, of the MCYTS and Project E-64-FEDER, D.G. Aragón.
Natalia Romero has been has been partially supported by the Spanish Ministry of Economy and Competitiveness, Project MTM2014-52016-C2-1-P.

References

[1] X. Chen and W. Chu, *Moments on Catalan number*, J. Math. Anal. Appl., **349**, (2) (2009), 311–316.

[2] P.J. Miana and N. Romero, *Moments of combinatorial and Catalan numbers*, J. Number Theory, **130**(8) (2010), 1876–1887.

[3] P.J. Miana, H. Ohtsuka and N. Romero, *Sums of powers of Catalan triangle numbers*, Arxiv 1602.04347, (2016) 1-18.

[4] L. W. Shapiro, *A Catalan triangle*, Discrete Math. **14** (1976), 83–90.

[5] N. Sloane, *http://www.research.att.com/.*

[6] R. P. Stanley, *Catalan Numbers,* Cambridge University Press, 2015.

[7] H. Wilf and D. Zeilberger, *Rational functions certify combinatorial*, J. Amer. Math. Soc. **3** (1990), 147–158.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Evaluation of an Evolutionary Multi-objective Optimization algorithm on a ARM+GPU system

Juan José Moreno[1], Gloria Ortega[1], Ernestas Filatovas[2], José Antonio Martínez[1] and Ester M. Garzón[1]

[1] *Group of Supercomputation-Algorithms, Dpt. of Informatics, ceiA3, Univ. of Almería, 04120, Almería, Spain*

[2] *Inst. of Informatics and Mathematics, Vilnius Univ., Akademijos str. 4, LT-08663, Vilnius, Lithuania*

emails: jrm069@inlumine.ual.es, gloriaortega@ual.es, ernest.filatov@gmail.com, jmartine@ual.es, gmartin@ual.es

Abstract

Nowadays, the application of Evolutionary Multi-objective Optimization (EMO) algorithms in real-time systems receives considerable interest. In this context, the energy efficiency of computational systems is of paramount relevance. Recently, the use of embedded systems based on heterogeneous (CPU+GPU) platforms is consistently increasing. For example, NVIDIA Jetson cards are low-power computers designed for development of embedded applications. They incorporate Tegra processors which feature a CUDA-capable GPU. This way, Jetson cards can be considered as a prototype of low-power computer of High Performance Computing. In this work, our interest is focused on the NSGA-II algorithm, a well-known representative of EMO algorithms. Our purpose on the low-power computers is twofold: to determinate the size range of NSGA-II problems which can be solved and to evaluate the energy efficiency of different NSGA-II versions. The analysis of the results shows the use of the Jetson platform as a low consumption platform which allows to accelerate the execution of instances of the state-of-the-art EMO algorithm – NSGA-II.

Key words: evolutionary multiobjective algorithms, energy efficiency, low-power platform, Jetson, NSGA-II

  ISBN: 978-84-608-6082-2

1 Introduction

Nowadays, energy costs represent a relevant share of the total costs of High Performance Computing (HPC) systems and they have become in a main issue. HPC platforms include several kinds of processing units, such as CPUs and GPUs, whose energy consumption depends on the kind of processing which is being performed. Lately, embedded hardwares have appeared thanks to the fast technological evolution. In this kind of systems, i.e. NVIDIA Jetson platforms, which combine low consumption and hybrid systems (CPU and GPU), the study of the energy consumption is relevant [13, 17]. Jetson is a low-cost and embedded platform widely used in real-time systems[1].

There are several real-time systems in which optimization problems are solved by Evolutionary Multi-objective Optimization (EMO) algorithms [2, 11]. EMO algorithms aim at finding an approximation of the Pareto set in a reasonable time. When this kind of algorithms work with relatively small populations, the use of NVIDIA Jetson platforms can be useful for two main reasons: (1) to accelerate the computation of EMO because of the exploitation of the multicore and the GPU which compose the NVIDIA Jetson; and (2) to decrease the energy-consumption because of the low consumption of these platforms.

Bearing in mind the limited computational resources on the Jetson platform and the characteristics of EMO problems, it is relevant to identify the scale of the problem that can be computed on this kind of platforms. Therefore, the idea behind this is as follows: 1) to determine the problem sizes that can be solved; and 2) to evaluate the energy efficiency of the multicore and/or the GPU versions of the EMO algorithms on the Jetson, according to the combination platform-resources and problem-size. We have focused on a state-of-the-art EMO algorithm – NSGA-II [4].

2 Background of the NSGA-II Algorithm

Many real-world problems are multi-objective, where several conflicting objective functions have to be optimized. The main aim of Multi-Objective Optimization (MOO) is to provide the set of solutions that determine the Pareto front used by the Decision Maker (DM). The majority of MOO problems are NP-hard and algorithms to approximate the Pareto front are widely-used. Evolutionary Multi-objective Optimization (EMO) approaches are commonly employed for this task [3, 16]. The set of obtained solutions is presented to the DM, who finally chooses one among them, according to his/her preferences.

Usually, a multi-objective algorithm can be organized in several stages [9]: evaluation of an objective function, Pareto dominance ranking (non-dominated sorting) and genetic operations. Examples of EMO approaches based on Pareto dominance ranking are: PESA-II [1], NSGA-II [4], R-NSGA-II [5], Synchronous R-NSGA-II [6], MOGA [7], PAES [8],

[1] http://www.nvidia.com/object/jetson-tk1-embedded-dev-kit.html

NSGA [15], SPEA2 [18], etc. The most computationally expensive part of those stages is the dominance ranking. The fast non-dominated sorting (FNDS) procedure, which is proposed and implemented in NSGA-II, has a complexity order of $O(MN^2)$, where M is the number of objective functions and N is the number of individuals [4]. As shown in [14], non-dominated sorting procedure consumes most of the computational burden of the EMO algorithm. However, only few attempts have been done in order to implement the Pareto dominance in parallel [12]. Consequently, to speed-up the non-dominated sorting procedure by using parallel computing is currently an appealing research line and can reduce the wall-clock time of EMO algorithms.

In this research, we focus on the energy efficiency evaluation of the most time consuming procedure – non-dominated sorting –, which is used in the state-of-the-art multi-objective genetic algorithms.

3 Energy Efficiency Evaluation

The performance per watt (ratio computational power over electrical power) plays a key role for evaluating the efficiency of the systems in terms of performance and power/energy [10]. Its increments mean that the system achieves better performance with less electrical power and also less energy consumed by the system. Therefore, it can be used as indicator of the energy efficiency of the system in terms of energy and performance.

The hardware setup we use is based on a NVIDIA Jetson TK1 development board[2], embedding a Tegra K1 SoC (Systems on a Chip) processor with 2GB of DDR3L RAM, and an energy meter (Watts up [3]).

On the one hand, the Tegra K1 includes an NVIDIA GPU with 192 CUDA Kepler cores and an ARM quad-core Cortex-A15 variant of low power architecture at 2.32 GHz. On the other hand, Watts up is an advanced plug load energy meter. With its simple operation, one can accurate monitor different plug load to determine their power consumption.

Jetson platform has been selected for three main reasons: (1) its SoC contains both a multicore-CPU as well as a GPU; (2) the independent control of the CPU and GPU frequencies is possible; (3) low power systems are gaining popularity and they are promising tools to build blocks of evolutionary algorithms for HPC platforms.

The memory requirements of the EMO problems depends on three parameters: the size of the population, the number of variables and the number of objective functions. These parameters have to be defined bearing in mind that Jetson board has only 2GB of physical memory, that is shared by the ARM CPU and the CUDA GPU. Therefore, we have considered instances of the NSGA-II problem which fit on the Jetson platform.

[2]http://www.nvidia.com/object/jetson-tk1-embedded-dev-kit.htm
[3]https://www.wattsupmeters.com/secure/index.php

In this work, the performance of the unified memory structure of the TK1 and also the power-performance ratio, as well as energy use, when executing an EMO algorithm on a Jetson platform has been evaluated.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science throughout projects TIN2012-37483, TIN15-66680, FIS-2015-69022-P and CAPAP-H5 network TIN2014-53522, by J. Andalucía through projects P12-TIC-301 and P11-TIC7176, and by the European Regional Development Fund (ERDF). Ernestas Filatovas has been partially granted by the European COST Action IC1305: Network for sustainable Ultrascale computing (NESUS).

References

[1] D.W. Corne, N.R. Jerram, J.D. Knowles, and M.J. Oates. PESA-II: Region-based selection in evolutionary multiobjective optimization. In *GECCO*, pages 283–290, 2001.

[2] C.E. Cortés, D. Sáez, F. Milla, A. Nuñez, and M. Riquelme. Hybrid predictive control for real-time optimization of public transport systems' operations based on evolutionary multi-objective optimization. *Transportation Research Part C: Emerging Technologies*, 18(5):757 – 769, 2010.

[3] K. Deb. *Multi-objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, 2001.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE T. Evolut. Comput.*, 6(2):182–197, 2002.

[5] K. Deb, J. Sundar, N. Udaya Bhaskara Rao, and S. Chaudhuri. Reference point based multi-objective optimization using evolutionary algorithms. *International Journal of Computational Intelligence Research*, 2(3):273–286, 2006.

[6] E. Filatovas, O. Kurasova, and K. Sindhya. Reference point based multi-objective optimization using evolutionary algorithms. *Informatica*, 26(1):33–50, 2015.

[7] C.M. Fonseca et al. Genetic algorithms for multiobjective optimization: Formulationdiscussion and generalization. In *ICGA*, volume 93, pages 416–423, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[8] J.D. Knowles and D.W. Corne. Approximating the non-dominated front using the Pareto archived evolution strategy. *Evol. Comput.*, 8(2):149–172, 2000.

[9] A. Lančinskas and J. Żilinskas. Solution of multi-objective competitive facility location problems using parallel NSGA-II on large scale computing systems. In *PARA*, pages 422–433. Springer, 2013.

[10] J. Leng et al. GPUWattch: Enabling Energy Optimizations in GPGPUs. *SIGARCH Comput. Archit. News*, 41(3):487–498, June 2013.

[11] P. Lokuciejewski and P. Marwedel. *Worst-case execution time aware compilation techniques for real-time systems*. Embedded systems. Springer, Dordrecht, Heidelberg, New York, 2011.

[12] G. Ortega, E. Filatovas, and E.M. Garzón. Parallelization of the non-dominated sorting procedure. In *Proceedings of 27th European Conference on Operational Research EURO 2015*, July 12–15 2015.

[13] J.P. Silva, E. Dufrechou, P. Ezzati, E. Quintana-Ortí, A. Remón, and P. Benner. Balancing energy and performance in dense linear system solvers for hybrid ARM+GPU platforms. *CLEI 2015 special issue - part 1*, 19(1):1–13, April 2016.

[14] C. Smutnicki, J. Rudy, and D. Żelazny. Very fast non-dominated sorting. *Decision Making in Manufacturing and Services*, 8(1-2):13–23, 2014.

[15] N. Srinivas and K. Deb. Multiobjective optimization using non-dominated sorting in genetic algorithms. *Evol. Comput.*, 2(3):221–248, 1994.

[16] E. Talbi. *Metaheuristics: from Design to Implementation*, volume 74. John Wiley & Sons, 2009.

[17] Y. Ukidave, D. Kaeli, U. Gupta, and K. Keville. Performance of the NVIDIA Jetson TK1 in HPC. In *2015 IEEE International Conference on Cluster Computing*, pages 533–534, Sept 2015.

[18] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. Technical Report 103, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Zurich, Switzerland, 2001.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Time series representation using fuzzy logic

Antonio Moreno-Garcia[1], Juan Moreno-Garcia[1], Luis Jimenez[1] and Luis Rodriguez-Benitez[1]

[1] *Department of Information and System Technologies, Universidad de Castilla-La Mancha*

emails: antmorgarcia@gmail.com, juan.moreno@uclm.es, luis.jimenez@uclm.es, luis.rodriguez@uclm.es

Abstract

The aim of this paper is to represent time series in a fuzzy way by means of a piecewise linear segment method. This technique allows representing the time series in a simple and efficient way. Furthermore, such representation collects the uncertainty generated in the process of generation of the segments. In order to obtain this representation, two stages are needed. First of all, we obtain a representation of the time series based on segments and commonly named as piecewise linear segment. After that, this is converted to the fuzzy domain in order to be used in a great range of applications. Several examples are shown in order to know how an input time series can be properly represented.

Key words: time series, fuzzy representation, database queries.

1 Introduction

Time series (TS) are used in a great number of applications, for instance, image processing [1], economics [2, 3], social sciences [4, 5] or sports [6]. Raw data are taken from sensors and data capture systems, then, they are stored and lastly they can be queried. Nowadays, there exists an important effort in research, development and innovation in all this process. Time series usually represents information as "raw data", that is, a set of values taken continuously from a device or system in constants intervals of time. In a formal way, they can be represented like shown in Equation 1.

$$Y = \{y_1, y_2, \ldots, y_n\} \tag{1}$$

 ISBN: 978-84-608-6082-2

where y_i is a value of the TS at the instant i and $1 \leq i \leq n$.

This way of representation of the data presents several problems. Maybe, the most important one is the necessity of great amounts of memory in order to be stored. This problem is even more pronounced if instead of considering an unique sensor we use different capture devices. Managing such great amount of information is a complex problem that must be addressed. For this reason, it is interesting to develop new types of representations in order to reduce memory consumption allowing to perform operations over data in a more efficient way. A technique used is named piecewise linear segment and it consists of representing the time series by means of a set of segments. Each segments corresponds to a piece of the time series.

A review of related research is presented in [7, 8] and a more recent one can be found in [9]. First proposal to segment series using a set of mathematical functions was done by Shatkay [10, 11]. Lastly Keogh et al. [7] designed the algorithm named SWAB (*Sliding Window and Bottom-up*). This algorithm uses a Bottom-Up method combined with a sliding window. Fuchs et al. [12] created a technique known as *SwiftSeg* by means of a polynomial approximation of TS taking as basis orthogonal polynomials in an sliding window. Furthermore, Huang et al. [9] used interpolation as tool and Garcia-Treviño et al. [13] presented a framework of representation of TS that uses dependence among the data of the own series together with algorithms of statistical classification. This technique of representation is very efficient and fast with respect to execution time although it presents a problem of an increasing error due to the raw data transformation to the set of segments obtained. Furthermore, it presents the error derived from the noise in the data capture. Fuzzy logic allows managing these problems because of its characteristics related to the data capture and after that the transformation into segments.

Kacprzyk et al. [14] introduced a method to create linguistic summaries from Time Series. This interesting paper details a proposal consisting of next stages:

1. Trend generation: it computes the segments with a concrete technique based on the algorithm of Sklansky et al. [15].

2. Representation of the characteristics of the TS: they present a methodology based on fuzzy logic and used later in our paper.

3. Generation of linguistic summaries: they use protoforms and calculus of linguistically quantified propositions. It is the most relevant part of their proposal although it is beyond the scope of this paper.

Our paper focuses in stage two, where the segments are represented by using fuzzy logic. Kacprzyk et al. consider three main aspects of their paper:

- Dynamics of change: for them it is the speed of change and to characterize it they use the slope of a line representing the trend. They define fuzzy membership functions to obtain a fuzzy granularity. For instance, they classify the trend like *quickly decreasing*, *decreasing*, *slowly decreasing*, *constant*, *slowly increasing*, *increasing* and *quickly increasing*.

- Duration: described like the length of the trend. They use a linguistic variable in order to represent it.

- Variability: it refers to how spread out (vertically, in the sense of values taken on) a group of data. So, they try to know if the trend has been well characterized with the segment or there is too much noise or uncertainty.

The proposal presented in this paper is different. We model the segments in a fuzzy way and then we can define operations between them. The use of fuzzy numbers as output value from an input value is proposed. This numbers are obtained in an automatic way from the segments taking into account an error measure that considers the original values of the time series in the domain the segment overlaps. The aspects consider by Kacprzyk can be then directly obtained from the own fuzzy segments. Anyway, duration is not considered although it could be computed identifying the initial and final times available in our proposal. Referring to the dynamics it is computed using the angle that defines the segment. On the other hand, the concept of fuzzy segment was introduced by Hoover in [16]. These segments were originally designed to detect the convergence lines in the shapes of an image. More concretely, they tried to automatically locate the optical nerve in an ocular fundus image. The generation of a fuzzy segment is based on the assumption that line-like shapes only contribute to a perception of a convergence only in their near neighbourhood. With this model, they develop a voting method to determine the convergence image. Their results in twenty images obtain a success rate of a 65%.

The rest of this paper is organised as follows: Section 2 details main part of our proposal. After that, in Section 3, we present a detailed example and the final results of several tests. Finally, Section 3.1 contains a discussion about the method and the results obtained and in Section 4 conclusions and future works are described.

2 Fuzzy piecewise linear segment

Before detailing our proposal it is interesting to know that the first step of it consists of the conversion of the TS into a set of ordered segments. There are several methods to transform the data but we have selected one designed by our own research group that is based on a sliding window with a low computational cost. This algorithm obtains as output a set of segments that it's going to be the input of the proposal introduced in this paper. The set

of segments is represented in Equation 2.

$$S = \{s_{f_1,l_1}, s_{f_2,l_2}, \ldots, s_{f_m,l_m}\} \tag{2}$$

where each segment $s_{f_k,l_k} = (m_{f_k,l_k} * x) + c_{f_k,l_k}$ is defined in a time interval delimited by first instant f_k and the last instant l_k.

A segment s_{f_k,l_k} is formally represented by using a 2-tuple: $s_{f_k,l_k} = \{m_{f_k,l_k}, c_{f_k,l_k}\}$.

As our goal is to obtain a representation based on a set of fuzzy segments, this means that taking as input a $t_i \in R1$ it is obtained a fuzzy triangular number as output named fn_i where t_i: $fs(t_i) = fn_i$, $t \in R1$ and fn_i is a fuzzy number.

Our representation proposal is named "Fuzzy Piecewise Linear Segment" (FPLS). A $FPLS$ is formally represented by means of Equation 3.

$$FPLS(T) = \{fpls_{f_0,l_0}, fpls_{f_1,l_1}, \ldots fpls_{f_{|fpls|},l_{|fpls|}}\} \tag{3}$$

where each $fpls_{f_k,l_k} = \{m_{f_k,l_k}, c_{f_k,l_k}, p_{f_k,l_k}\}$ and p_{f_k,l_k} is the average of ratio error defined in Equation 5.

A fuzzy segment $fpls_{f_k,l_k}$ is modelled using a 3-tuple $\{m_{f_k,l_k}, c_{f_k,l_k}, p_{f_k,l_k}\}$ containing the slope m_{f_k,l_k}, the intercept c_{f_k,l_k} and the average of ratio error p_{f_k,l_k} of a segment. Using this three values the output fuzzy set fn_i can be computed taken as input value a t_i. Equation 4 shows how $fpls_{f_k,l_k}$ can generate the triangular fuzzy number fn_i as output with a support obtained from the values fn_i^a, fn_i^b and fn_i^c.

$$fpls_{f_k,l_k} = \begin{cases} if\ t_i < f_k: \quad without\ output \\ if\ f_k \leq t_i \leq l_k: \quad \{fn_i^a, fn_i^b, fn_i^c\} = compute(fpls_{f_k,l_k}, t_i) \\ if\ l_k < t_i: \quad without\ output \end{cases} \tag{4}$$

where $compute(fpls_{f_k,l_k}, t_i)$ is a function calculating fn_i^{DOWN}, fn_i^s and fn_i^{UP}; and finally it obtains a triangular fuzzy set as output: $fn_i = \{fn_i^a, fn_i^b, fn_i^c\}$.

For values of t_i less than f_k there is no output because the segment is not defined for these values of the domain. It's the same for values greater than l_k. But, for values between f_k and l_k it is computed by using two parallel segments with respect to s_{f_k,l_k} named like UP_{f_k,l_k} and $DOWN_{f_k,l_k}$ (Figure 1). An error measure is used too and it is called "average of the error ratio" (Equation 5). This equation computes the average of the error ratio for every segment.

$$p_{f,l} = \frac{\sum_{i=f}^{l} \frac{|s_{f,l}(t_i) - y_i|}{y_i}}{l - f + 1} \tag{5}$$

where f and l are the instants for the beginning and end of the segment and $s_{f,l}(t_i)$ is the value of the segment $s_{f,l} \in S$ in the instant t_i.

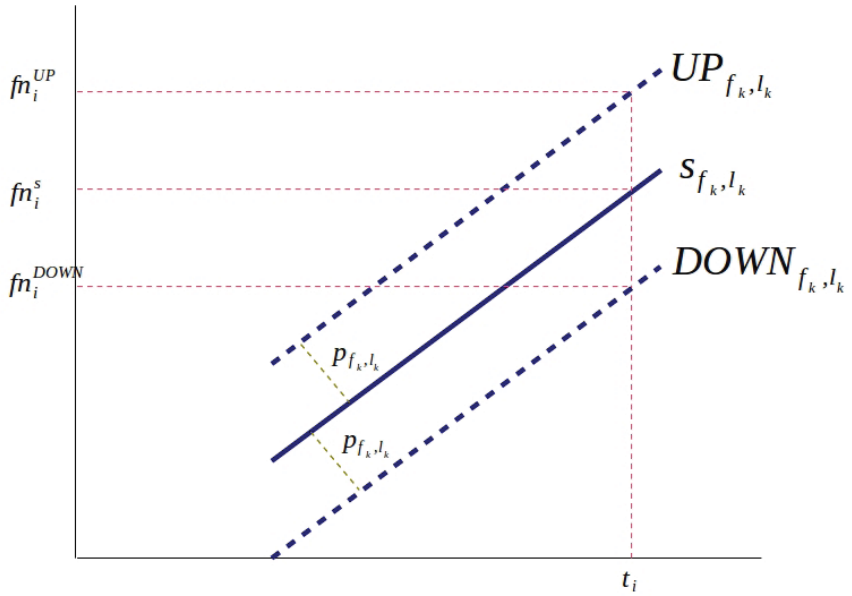


Figure 1: Computation of fn_i.

p_{f_k,l_k} is needed to compute the segments UP_{f_k,l_k} and $DOWN_{f_k,l_k}$ are up and down in the y-axis with respect to s_{f_k,l_k}, that is:

- $UP_{f_k,l_k} = (m_{f_k,l_k} * x) + c_{f_k,l_k} + p_{f_k,l_k}$.
- $DOWN_{f_k,l_k} = (m_{f_k,l_k} * x) + c_{f_k,l_k} - p_{f_k,l_k}$.

Values fn_i^a, fn_i^b and fn_i^c defining fn_i are computed as the output value UP_{f_k,l_k} ($fn_i^{UP} = UP_{f_k,l_k}(t_i)$), fn_{f_k,l_k} ($fn_i^s = s_{f_k,l_k}(t_i)$) and $DOWN_{f_k,l_k}$ ($fn_i^{DOWN} = DOWN_{f_k,l_k}(t_i)$) for the value t_i and in increasing order.

As the support of fn_i is computed depending on the average of the error ratio from the obtained segment, the average for the uncertainty is appropriate. That is, the greater the error is the greater the support of the fuzzy number is.

This way of representation allows comparing the fuzzy number with crisp values, other fuzzy numbers or whatever other representation taking advantage of the powerful operations of the fuzzy logic and a proper managing of uncertainty.

3 Experimental results

In this section, an example is introduced where we show how the proposed approach models a TS, detailing some properties of the obtained results. In addition, some ideas related to

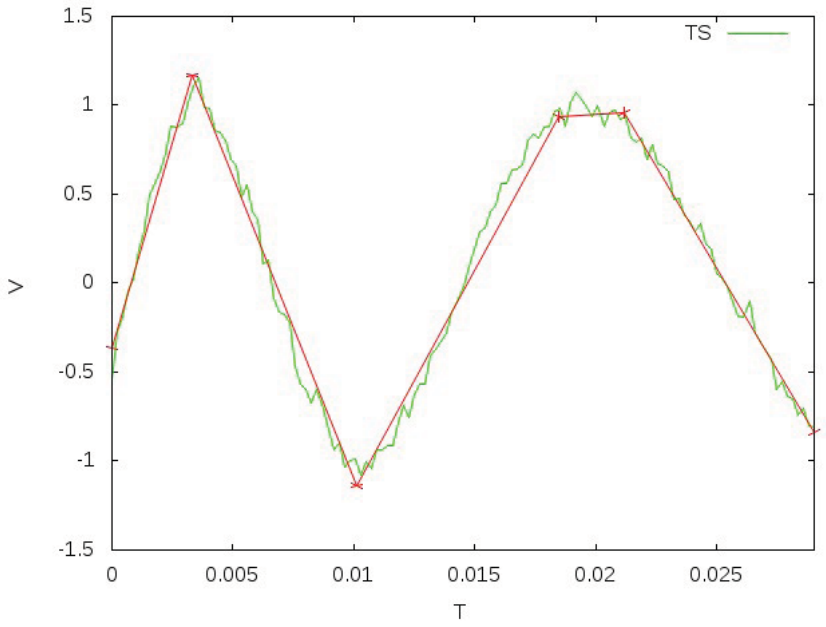


Figure 2: Set S obtained as result.

potentials fields of application are presented.

First of all, it is needed a set S obtained using a method able to generate the segments from the TS. Figure 2 shows graphically the TS taken as input and the segments positioning. It can be observed how the five segments are well adjusted with respect to data.

S and the original TS have been used as input in order to obtain the set of fuzzy segments $FPLS$. Table 1 details the obtained set T. First column represent the identifier of the segment with the initial and final instants (f_k and l_k), while the three last rows correspond to the slope, intercept and the average of ratio error for each segment.

In order to check how the resulting set of fuzzy segments represents the output of the TS, the output fuzzy numbers have been generated for a set of values from TS equally distanced (Table 2). First and second cols show t_i and y_i from every part selected from the input TS. Third column shows the fuzzy number obtained fs_i while the fourth col shows the membership value for v_i to the fuzzy number fs_i that demonstrates how is modelled each fs_i to its value represented by TS y_i. Finally, last column shows position k of $fpls_{f_k,l_k} \in T$ using to obtain the output fuzzy number fs_i.

As it can be observed, every input value is related to a output fuzzy number. The obtained $FPLS$ can be used in different applications. For example, to establish comparisons between different TSs. Now, it is going to be described a first approach to the comparison

Table 1: Set T for the example.

$fpls_{g_k,f_k}$	m_{f_k,g_k}	c_{f_k,g_k}	p_{f_k,g_k}
$fpls_{0.0,\ 0.0033}$	458.9375	-0.3616	0.2574
$fpls_{0.0033,\ 0.0101}$	-339.4396	2.3022	0.2144
$fpls_{0.0101,\ 0.0185}$	248.9599	-3.6608	0.2433
$fpls_{0.0185,\ 0.0212}$	6.9336	0.8103	0.0517
$fpls_{0.0212,\ 0.0290}$	-228.6086	5.79461	0.1421

Table 2: Resulting fuzzy numbers.

t_i	y_i	fs_i	$\mu_{fpls_{f_k,g_k}}(v_i)$	k
0.0000	-0.5424	$[-0.619, -0.3616, -0.1042]$	0.2975	
0.0013	0.2897	$[-0.0043, 0.2531, 0.5105]$	0.8577	1
0.0027	0.8796	$[0.6104, 0.8678, 1.1252]$	0.954	
0.004	0.9803	$[0.724, 0.9384, 1.1527]$	0.8047	
0.0054	0.4847	$[0.2694, 0.4837, 0.6981]$	0.9956	
0.0067	-0.0879	$[-0.1853, 0.0291, 0.2435]$	0.4541	2
0.008	-0.5996	$[-0.6399, -0.4255, -0.2111]$	0.1878	
0.0094	-0.8971	$[-1.0945, -0.8801, -0.6657]$	0.9209	
0.0107	-1.0406	$[-1.2366, -0.9933, -0.75]$	0.8055	
0.0121	-0.6873	$[-0.9032, -0.6599, -0.4166]$	0.8872	
0.0134	-0.3716	$[-0.5697, -0.3264, -0.0832]$	0.8143	3
0.0147	0.0908	$[-0.2363, 0.007, 0.2503]$	0.6556	
0.0161	0.5588	$[0.0972, 0.3404, 0.5837]$	0.1024	
0.0174	0.8367	$[0.4306, 0.6739, 0.9172]$	0.3309	
0.0188	0.8866	$[0.8886, 0.9404, 0.9921]$	0.0	4
0.0201	0.9952	$[0.8979, 0.9496, 1.0014]$	0.1196	
0.0214	0.8154	$[0.7535, 0.8957, 1.0378]$	0.4355	
0.0228	0.6605	$[0.4473, 0.5895, 0.7316]$	0.5001	
0.0241	0.2898	$[0.1412, 0.2833, 0.4254]$	0.9545	5
0.0254	-0.017	$[-0.165, -0.0229, 0.1193]$	0.9585	
0.0268	-0.331	$[-0.4712, -0.3291, -0.1869]$	0.9864	
0.0281	-0.6484	$[-0.7774, -0.6353, -0.4931]$	0.9081	

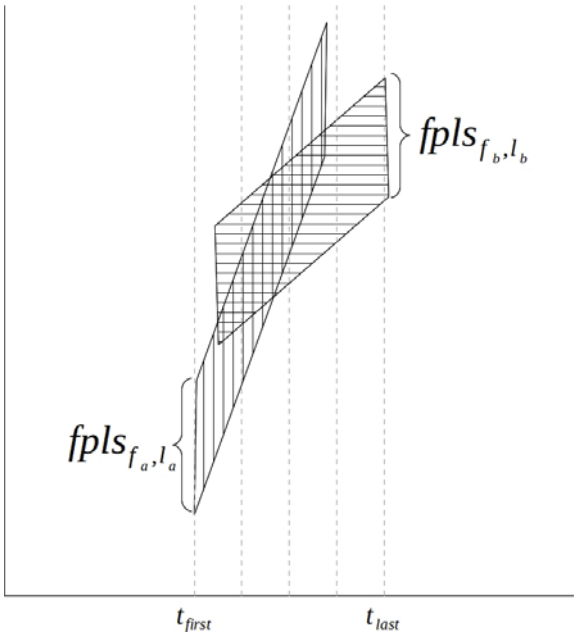


Figure 3: Comparison between to fuzzy segments.

between segments. The main idea of the comparison would be similar to the algorithms proposed in [17, 18]. Figure 3 shows the comparison process of $fpls_{f_a,l_a}$ and $fpls_{f_b,l_b}$. It must be compared the output fuzzy numbers for the values belonging to the time interval overlapped by both segments ($[t_{first}, t_{last}]$) equally distanced (in the figure they are represented by means of vertical dotted line. There are several methods to compare fuzzy number that can be used in order to do this.

Finally, we must remark that the proposed representation of the segments can be used to represent the TS in a similar way as the one proposed by Kacprzyk.

3.1 Discussion

We can affirm that the most similar paper to ours is the one of Kacprzyk et al. [14] where they propose to obtain fuzzy sets in order to represent the complete segment with the aim of having a global vision of it. Lastly, it is used to produce a description of the TS. Kacprzyk et al. present interesting concepts such as dynamics of change and duration (obtaining a measure by means of fuzzy sets), and the variability of the segment using only statistical measures. Our goal is to represent each segment to obtain information from every instant of time. That's why the concept of fuzzy segment returning a fuzzy number for every input value is used.

Using fuzzy numbers give the possibility of capturing the uncertainty of the representation using segments, due to the error and to offer a range of techniques to compare directly the resulting fuzzy segments with the TS, with linear segments or another fuzzy segments.

Then, we believe this is a powerful representation that can be used in a wide range of applications which need to obtain high level information from the TS. A concrete example can be queries of descriptions of TSs.

4 Conclusions and future works

This paper presents a new way of representation for TS based on fuzzy logic. It represent the time series as a "fuzzy piecewise linear segment" based on the definition of the fuzzy segment. Each fuzzy segment allows obtaining a fuzzy number as output for every time instant of input. Furthermore, some initial ideas of the comparison between fuzzy segments are introduced. We consider this representation very adequate to be used in different applications that works with TSs as linguistic descriptions or advanced queries over TSs. A brief approach of new lines of research has been presented too.

Acknowledgements

This work has been funded by the project TIN2015-64776-C3-3-R of the Spanish Government.

References

[1] Romani, L., De Avila, A., Chino, D., Zullo, J., Chbeir, R., Traina, C., Traina, A.: A new time series mining approach applied to multitemporal remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing. 51(1) (2013) 140-150.

[2] Wright, M., Stern, P.: Forecasting new product trial with analogous series. Journal of Business Research. 68(8) (2015) 1732-1738.

[3] Wei., L-Y.: A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting. Applied Soft Computing, 42 (2016) 368-376.

[4] Liu, L., Peng, Y., Wang, S., Liu, M., Huang, Z.: Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors. Information Sciences, 340-341 (2016) 1-17.

[5] Sanchez-Valdes, D., Alvarez-Alvarez, A., Trivino, G.: Dynamic linguistic descriptions of time series applied to self-track the physical activity. Fuzzy Sets and Systems, 285 (2016) 162-181.

[6] Moreno-Garcia, J., Abián-Vicén, J., Jimenez-Linares, L., Rodriguez-Benitez, L.: Description of multivariate time series by means of trends characterization in the fuzzy domain. Fuzzy Sets and Systems, 285 (2016) 118-139.

[7] Keogh, E., Chu, S., Hart, D., Pazzani, M.: An Online Algorithm for Segmenting Time Series. Proc. IEEE Int'l Conf. Data Mining, (2001) 289-296.

[8] Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting Time Series: A Survey and Novel Approach. Data Mining in Time Series Databases, M. Last, A. Kandel, and H. Bunke, eds., World Scientific Publishing, 57, ch. 1, (2004) 1-22.

[9] Huang, H., Matija, M., Suykens, J.A.K.: Hinging Hyperplanes for Time-Series Segmentation. IEEE Trans. Neural Networks and Learning Systems, (2013) 24(8) 1279-1291.

[10] Shatkay, H.: Approximate Queries and Representation for Large Data Sequences. Technical Report CS-95-03, Brown University, February 1995.

[11] Shatkay, H., Zdonik, S.: Approximate queries and representations for large data sequences. Proc. 12th Int'l Conf. on Data Engineering, (1996) 536-545.

[12] Fuchs, E., Gruber, T., Nitschke, J., Sick, B.: Online Segmentation of Time Series Based on Polynomial Least-Squares Approximations. IEEE Pattern Analysis and Machine Intelligence. 32(12) 2232-2245.

[13] Garcia-Treviño, E.S., Barria, J.A.: Structural generative descriptions for time series classification. IEEE Transactions on Cybernetics. 44(10) (2014) 1978-1991.

[14] Kacprzyk, J., Wilbik, A.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets and Systems, 159 (2008) 1485-1499.

[15] Sklansky, J., Gonzalez, V.: Fast polygonal approximation of digitized curves. Pattern Recognition. 12 (1980) 327331.

[16] Hoover, A., Goldbaum, M. : Fuzzy convergence. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (1998) 716-721.

[17] Moreno-Garcia, J., Castro-Schez. J.J., Jimenez, L.: A New Method to Compare Dynamical Systems Systems. P. Melin et al. (Eds.), Springer-Verlag Berlin Heidelberg 2007, LNAI 4529. (2007) 533-542, 2007.

[18] Moreno-Garcia, J., Castro-Schez. J.J., Jimenez, L.: A New Method to Compare Dynamical Systems Modeled Using Temporal Fuzzy Models. Uncertainty and Intelligent information Systems, 22 (2008), 307-320.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Application of GFDM: Modelling of Geophysical methods

A. Muelas[1], E. Salete[1], J.J. Benito[1], F. Ureña[2], L. Gavete[3] and M. Ureña[1]

[1] *UNED, ETSII, Madrid, Spain*

[2] *UCLM, IMACI, Ciudad Real, Spain*

[3] *UPM, ETSIM, Madrid, Spain*

emails: amuelas@ind.uned.es, esalete@ind.uned.es, jbenito@ind.uned.es, fuprieto@terra.com, lu.gavete@upm.es, miguelurenya@gmail.com

Abstract

Geophysical methods can be used indirectly to map some geological features in detail, such as faults, shear zones, karst, alteration zones and other structures. Two widespread geophysical methods in civil engineering are: the Cross-hole method and the seismic refraction method. These geophysical investigation methods can be modelled by numerical methods.

This article shows the use of the Generalized Finite Difference Method to model the aforementioned geophysical methods and analyze the results obtained.

Key words: meshless methods, generalized finite difference method, geophysics

1 Introduction

During recent years, meshless methods have appeared as a type of numerical methods which are capable of overcome the difficulties arisen in conventional computational mesh based methods. The Generalized Finite Difference Method (GFDM) is based on the concept of using a N-node star and weighting functions, to obtain finite difference formulae for irregular meshes.

GFDM has proven to be a successful method to model seismic wave propagation using regular or irregular meshes [1]. A closely related application is the modelling of geophysical investigation methods that have been used to image the subsurface of the Earth and determine site geology and stratigraphy.

In this paper, the GFDM has been applied to model two widespread geophysical tests, namely the Cross hole method and the Seismic refraction method.

 ISBN: 978-84-608-6082-2

2 Fundamentals of the GFDM

This section describes the fundamentals of the GFDM, obtaining the explicit generalized difference schemes for seismic wave propagation.

The equation of motion for a perfectly elastic, homogeneous, isotropic medium in the domain $\Omega \subset \mathbb{R}^2$ is the following one:

$$(\lambda + G)U_{k,ik} + GU_{i,kk} = \rho \ddot{U}_i \tag{1}$$

where U_i are the components of the displacement, ρ is the density, λ and G are the Lamés elastic coefficients.

In addition, two types of boundary conditions are considered: Dirichlet boundary conditions and free surface.

On the free surface the following conditions are imposed:

$$\sigma_{ij} n_j = \left(\lambda u_{k,k}\delta_{ij} + G\left(u_{i,j} + u_{j,i}\right)\right) n_j = g_i(t) \tag{2}$$

Where $g_i(t)$ is the i-component of the stress applied on the boundary. If $g_i(t)$ is equal to zero there are no forces applied on the boundary.

An irregular cloud of points is generated in the domain $\Omega \cup \Gamma$, where Γ is the boundary of the domain, with the intention of obtaining linear expressions for the approximation of partial derivatives at all the points of the domain. Every node in the domain has an associated star assigned to it. The star is defined as the set of nodes surrounding the central node that will contribute to the field variable approximation at the central node.

This scheme uses the central-difference form for the time derivative.

$$\frac{\partial^2 u_i(x_0, y_0, n\Delta t)}{\partial t^2} = \frac{{}^{n+1}u_i^0 - 2\ {}^{n}u_i^0 + {}^{n-1}u_i^0}{(\Delta t)^2} \tag{3}$$

Following references [3], [4], [5], [7], and [2], the following expression shows the explicit finite difference formulae for the second spatial derivatives with second and fourth order approximation (p=2, 4) for the spatial derivatives:

$$\frac{\partial^2 U_i(x_0, y_0, n\Delta t)}{\partial x_j \partial x_k} = \left[u_{i,jk}^0\right]_{t=n\Delta t} = -m_{jk}^{0,p}\ {}^{n}u_i^0 + \sum_{l=1}^{N} m_{jk}^{l,p}\ {}^{n}u_i^l + \Theta\left[(h_i)^p\right] \tag{4}$$

where capital letters are used for exact values and small letters are used for approximated values. The superscript n represents the time step ($t = n\Delta t$), the superscripts 0 and l refer to the central node and the rest of nodes of the star, respectively. N is the number of nodes in the star. The value N adopted for the second order approximation is $N = 8$, and $N = 30$ for the fourth order approximation, whereas the rest of nodes of the star are selected by using the distance criteria:

$$h_i^l = x_i^l - x_i^0 \tag{5}$$

In eq. 4, the coefficients $m_{jk}^{0,p}$ multiply the approximate value of the function U at the central node ${}^n u_i^0$ of the star, whereas the coefficients $m_{jk}^{l,p}$ multiply the approximate value of the function U at the rest of nodes of the star ${}^n u_i^l$, to obtain the generalized finite difference explicit expressions for the space derivatives, corresponding to a time $t = n\Delta t$. In all these expression the cross-terms are equal.

After replacing in eq. 1 the explicit finite difference expressions obtained for the spatial derivatives and the time derivatives, the following explicit difference scheme is obtained:

$$\begin{aligned} {}^{n+1}u_i^0 = 2\ {}^n u_i^0 - {}^{n-1}u_i^0 + \\ + \frac{(\Delta t)^2}{\rho}\left[(\lambda+G)\left(-m_{ij}^{0,p}\ {}^n u_j^0 + \sum_{l=1}^{N} m_{ij}^{l,p}\ {}^n u_j^l\right) + \right. \\ \left. + G\left(-m_{jj}^{0,p}\ {}^n u_i^0 + \sum_{l=1}^{N} m_{jj}^{l,p}\ {}^n u_i^l\right)\right] + \Theta\left[(\Delta t)^2, (h_i)^p\right] \end{aligned} \tag{6}$$

The idea of using an eight node star and weighting functions, to obtain finite difference formulae for irregular meshes, was first put forward in [6] using moving least squares (MLS) interpolation, and an advanced version of the GFDM was given in [3]. Benito et al. [4] reported that the solution of the generalized finite difference method depends on the number of nodes in the cloud, the relative coordinates of the nodes with respect to the star node and on the weight function employed.

The imposition of Dirichlet-type boundary conditions is evident. Nevertheless, in case of free surface it is necessary to express the applied stress as a function of the displacements:

$$\left(\lambda u_{k,k}\delta_{ij} + G\left(u_{i,j} + u_{j,i}\right)\right) n_j = g_i(t) \tag{7}$$

The free surface condition is satisfied by adding a series of nodes (Neumann nodes) in the same amount as the boundary nodes, and imposing the displacements of the added nodes so that the zero stress condition is fulfilled on the boundary nodes.

The following expression shows the explicit finite difference formulae for the first spatial derivatives:

$$\frac{\partial u_i(x_0, y_0, n\Delta t)}{\partial x_j} = \left[u_{i,j}^0\right]_{t=n\Delta t} = -m_j^0\ {}^n u_i^0 + \sum_{l=1}^{N} m_j^l\ {}^n u_i^l \tag{8}$$

After substituting the first order derivatives that appear in the free surface condition by the finite difference expression above, the following system of $2n$ equations is obtained

 ISBN: 978-84-608-6082-2

$$\left\{\lambda\left(-m_k^s\ {}^n u_k^s+\sum_{l=1}^{N} m_k^l\ {}^n u_k^l\right)+ G\left[\left(-m_j^s\ {}^n u_i^s+\sum_{l=1}^{N} m_j^l\ {}^n u_i^l\right)+\left(-m_i^s\ {}^n u_j^s+\sum_{l=1}^{N} m_j^l\ {}^n u_j^l\right)\right]\right\}\cdot n_j = g_i(t) \quad (9)$$

where the number of star nodes is $N = n_{se}+n_{si}$, and the $2n$ unknowns are the displacements in the n added nodes. These unknowns appear in the summation. By solving this system, the function values u_i^{se} are obtained on the n added Neumann nodes at the time $t = n\Delta t$.

3 Geophysical investigation modelling

Geotechnical geophysical tools are routinely used to image the subsurface of the Earth and determine site geology, stratigraphy, and rock quality. Commonly employed geophysical methods include seismic refraction, seismic reflection, MASW, cross-hole seismic tomography, electrical resistivity, GPR (Ground Penetrating Radar), electromagnetics, gravity, etc.

The *Cross hole* test allows to measure the velocity of seismic waves between boreholes. The typical approach involves lowering a three component triaxial geophone receiver down one hole while lowering a source down an adjacent hole, firing the source at some prescribed depth interval. The source is capable of generating shear and compressional waves. The source and geophone are always at the same elevation inside the boreholes. The energy from each shot is measured at a single depth in each receiver hole. The travel times measured are then converted to velocities by dividing them into the distance between the holes. By repeating the test at increasing depths, a seismic velocity vertical profile can be obtained.

In the *Seismic Refraction method* a seismic source (a hammer hitting on a plate, an explosive, etc.) is used to generate compressional waves, which are measured by a series of evenly spaced geophones located on the ground surface. Two types of waves are generated: the P-wave (a compressional, longitudinal wave) and the S-wave (a shear, transverse wave). P-waves propagate at the highest velocity of any seismic waves and are therefore used to pick the first arrival of seismic waves that propagated through ground materials.

Some head waves enter a high velocity medium near the critical angle and travel in the high velocity medium nearly parallel to the interface between layers. Since seismic waves move faster in the high velocity medium than the upper, the wave refracted along that interface will overtake the direct wave at some distance from the source. This point at which the refracted wave overtakes the direct wave arrival is known as the critical distance, and is used to estimate the depth to the refracting surface (figure 1). The refracted wave is

first detected at all subsequent geophones, at least until it is in turn overtaken by a deeper, faster refraction.

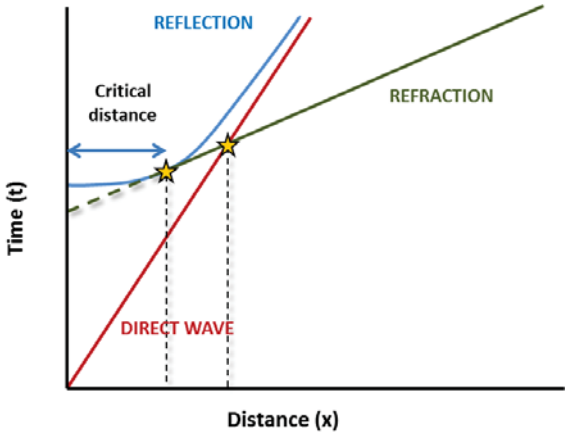


Figure 1: Seismic Refraction: Distance vs. Wave arrival time

Following, the Generalized Finite Difference Scheme has been applied to model two case studies of the aforementioned seismic tests.

3.1 Cross-hole test

A geotechnical investigation has been carried out to prepare a Construction Project in Saudi Arabia. Three boreholes with core recovery and an additional cross-hole test were performed in order to obtain a seismic velocity vs. depth profile.

Based on the geotechnical investigation aforementioned, the subsoil comprises an upper layer of medium dense sand up to 6 meters depth, overlying a very dense silty sand layer.

The dynamic Young's Modulus (E), the dynamic Shear Modulus (G) and the Lambda parameter (λ) are shown in figure 2. The lambda parameter is derived from the dynamic Young's Modulus or the dynamic Shear modulus by the expression given in eq. 10.

$$\lambda = \frac{E\nu}{(1-\nu)(1-2\nu)} = \frac{2\nu G}{1-2\nu} \tag{10}$$

The dynamic parameters shown in figure 2 have been assigned to the subsoil as a function of soil depth. In addition, the scheme shown in figure 3 has been used to model the cross-hole test.

In figure 4 it may be seen the moment in which the front shear wave reaches the geophone located at 9 meters depth.

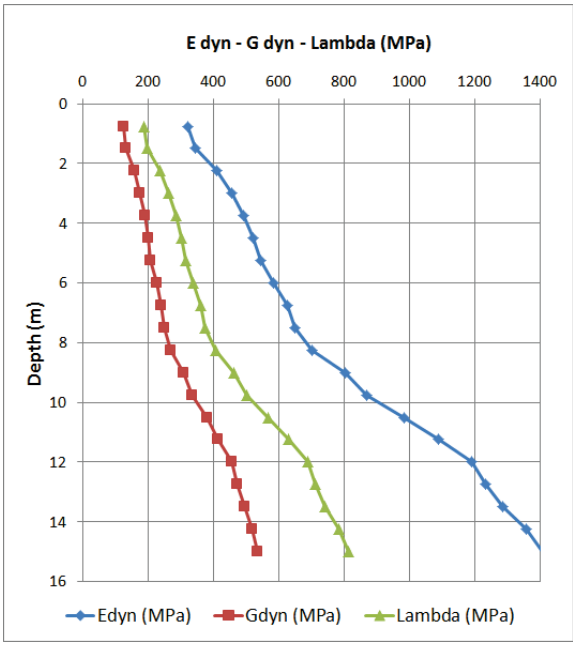


Figure 2: Dynamic Soil Properties (Borehole CH03)

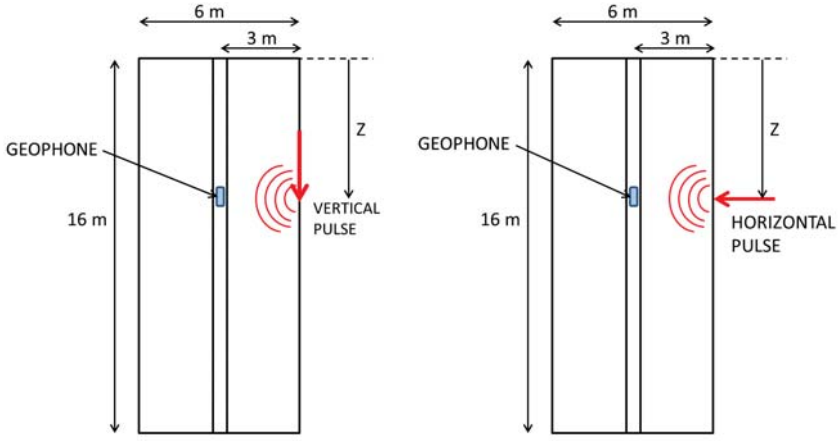


Figure 3: Scheme used to model the cross-hole test

 ISBN: 978-84-608-6082-2

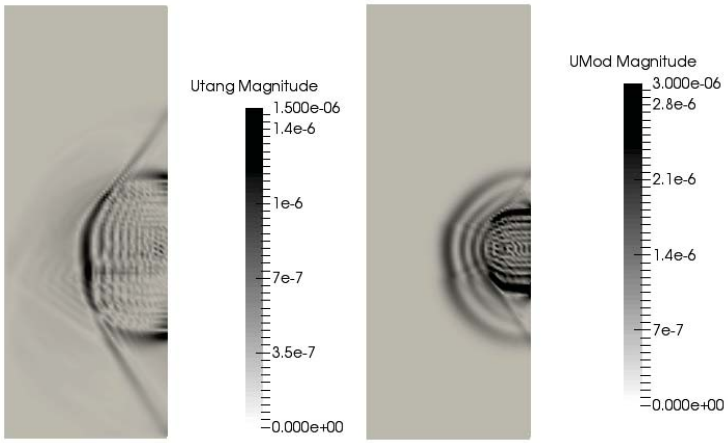


Figure 4: Front Shear waves (left) and P-wave (right) reaching the 9m-depth geophone

The displacement over time curve for the geophone located at 9 meters depth is shown in figure 5. From this curve, the shear wave velocity can be derived by dividing the distance between the source and the geophone by the time of wave arrival.

By applying a horizontal pulse at the source location, the P-waves propagation may be analyzed (figure 4).

The compression waves velocity (P-waves velocity) is derived by dividing the distance between the source and the geophone by the time of P-wave arrival.

Figure 6 shows the comparison between the actual seismic velocities obtained in the cross-hole test and those obtained in the model. As it may be observed from the chart, the model-calculated seismic velocities match satisfactorily to those obtained in the cross-hole test, with a 5% maximum deviation. It is therefore concluded that the model is capable to accurately reproduce a cross-hole test, and even obtain dynamic parameters (e.g. Dynamic Young's Modulus and Dynamic Shear Modulus) from cross-hole data.

3.2 Seismic Refraction Method

The seismic refraction method is usually performed in quarries or open-pit mines in order to make an estimate of the rock volume that may be extracted from the site. An example of the seismic refraction method carried out in a quarry in Valparaiso (Chile) has been modeled using the GFDM. As a result of the geophysical study, the longitudinal profile shown in figure 7 was obtained. Basically, the subsoil was composed of three layers with a seismic velocity that increases with depth.

In the GFDM model, it can be distinguished the refracted head waves that generate

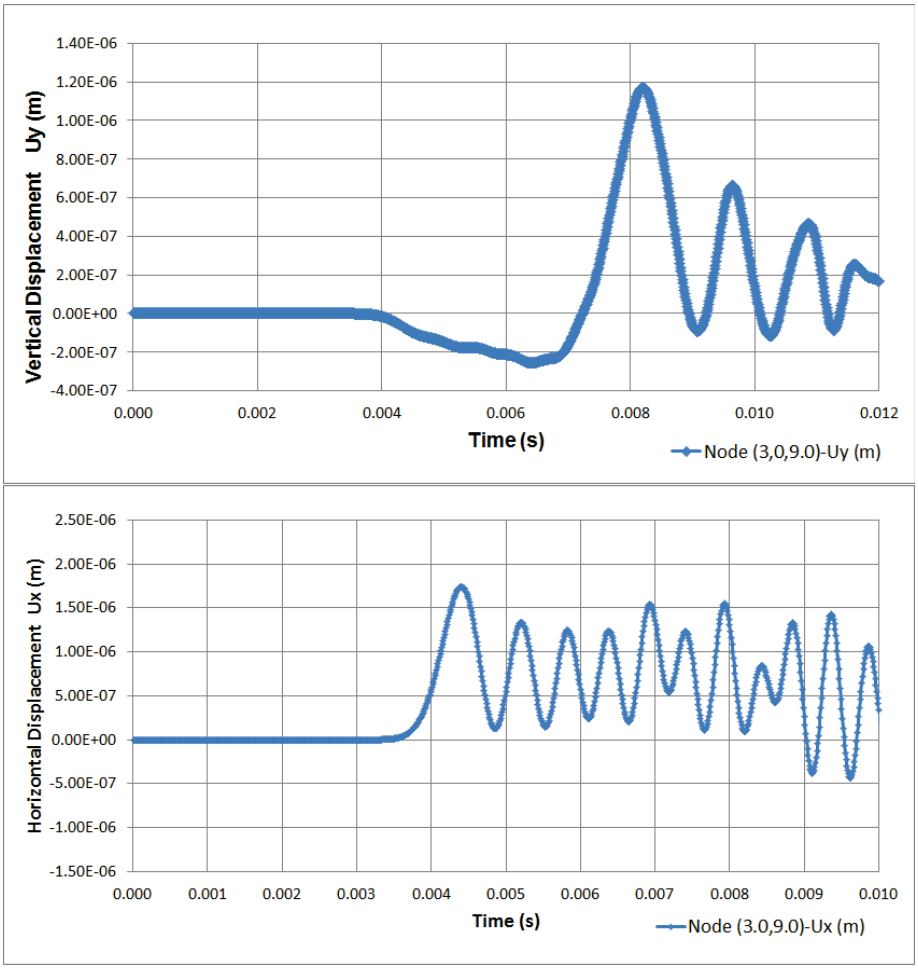


Figure 5: Vertical (up) and horizontal (down) displacement vs. time for geophone located at 9 m depth

between layers 1 and 2 (figure 8), towards the upper layer. The same figure clearly shows the refracted head waves between layers 1 and 2, and between layers 2 and 3, a step further.

The refracted head waves travel towards the upper layer (figure 9) when the waves reach the interface between both layers at a critical angle i_c, which is defined by expression 11.

$$\sin i_c = \frac{v_1}{v_2} \tag{11}$$

where v_1 and v_2 are the wave propagation velocity in the upper an lower layers respectively.

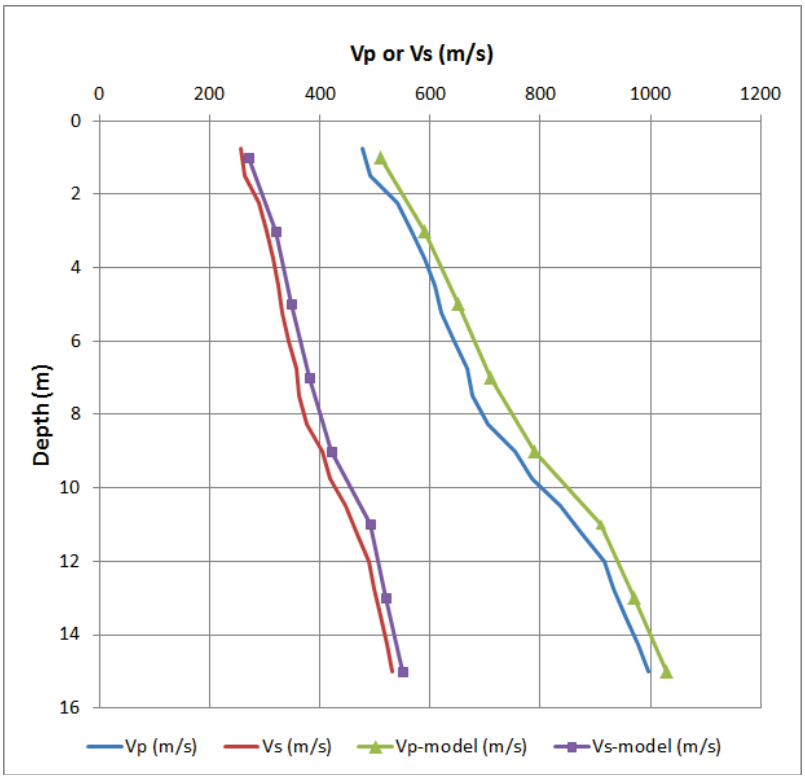


Figure 6: Comparison between actual and model-obtained seismic velocities

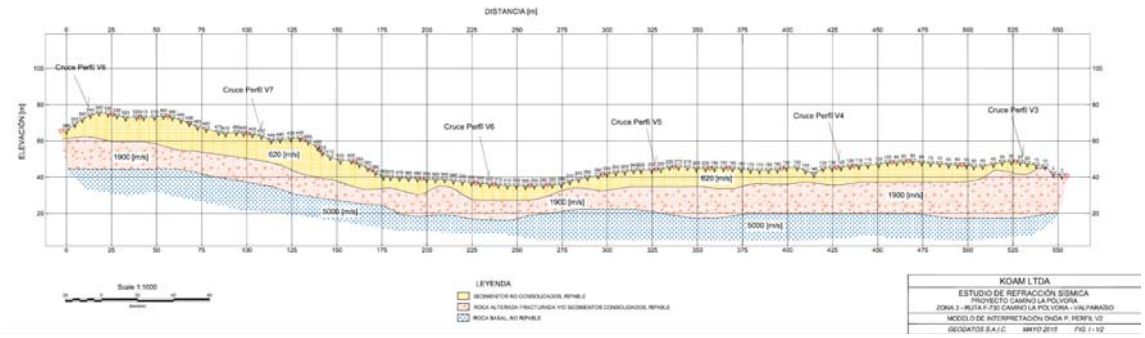


Figure 7: Longitudinal profile with seismic velocities

 ISBN: 978-84-608-6082-2

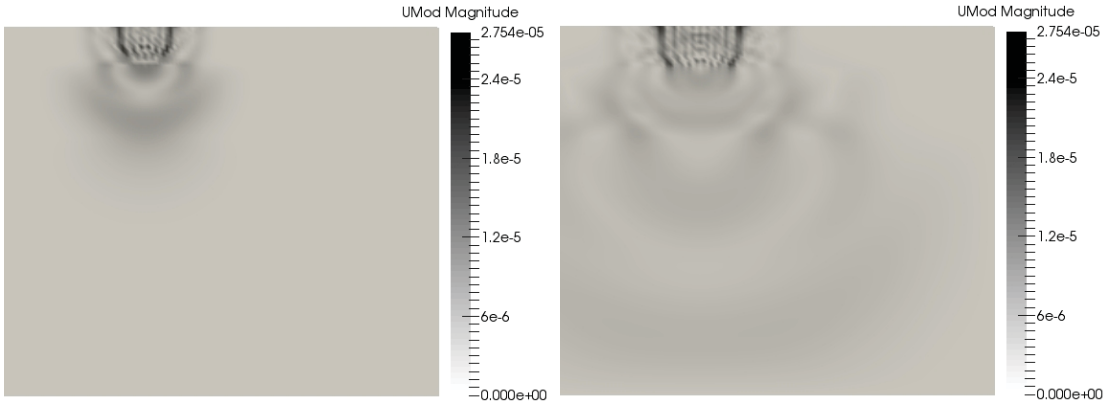


Figure 8: Refracted head waves between layers 1 and 2 (left) and layers 2 and 3 (right)

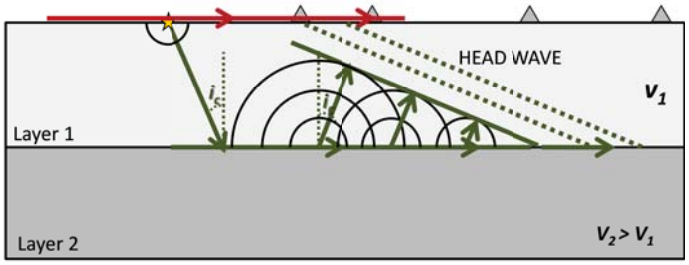


Figure 9: Critical angle to generate refracted waves

With the aim of checking the refracted head wave angle, a simpler model with two layers has been analyzed. This model is shown in figure 10.

The critical angle corresponding to the refracted waves is obtained from the following seismic velocities ratio:

$$\sin i_c = \frac{v_1}{v_2} = \frac{300\ m/s}{500\ m/s} = 0.6 \tag{12}$$

Therefore, the critical angle is $i_c = 37°$.

This angle perfectly matches with the refracted waves angle measured in the screenshots taken from the GFDM model, shown in figure 11. It is consequently considered that this GDFM model accurately reproduces the seismic refraction test.

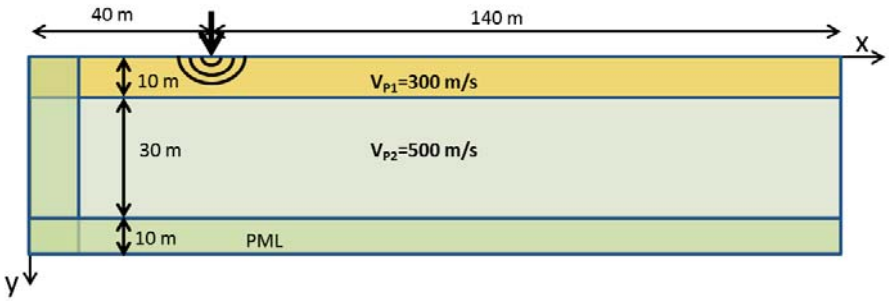


Figure 10: Model used to analyze the refracted waves angle

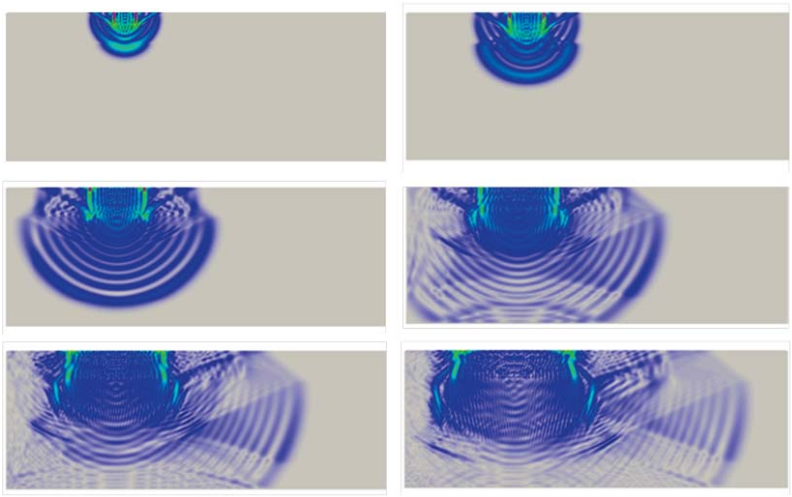


Figure 11: Seismic refraction screenshots

4 Conclusions

This paper describes the fundamentals of the GFDM, obtaining the explicit generalized difference schemes for seismic wave propagation.

It has been shown the use of the Generalized Finite Difference Method to model two of the most widespread geophysical methods used in Civil Engineering: the Cross-hole method and the Seismic Refraction method.

It is has been concluded that a GFDM model is capable to accurately reproduce a

cross-hole test, and even obtaining dynamic parameters (e.g. Dynamic Youngs Modulus and Dynamic Shear Modulus) from cross-hole data.

In addition, this article shows the use of the Generalized Finite Difference Method to accurately model the seismic refraction method, focusing on the seismic velocities and the angle of refracted waves.

Acknowledgement

The authors acknowledge the support of the *Escuela Técnica Superior de Ingenieros Industriales (UNED)* of Spain, project 2015-IFC02.

References

[1] E. Salete L. Gavete F. Ureña, J.J. Benito. A note on the application of the generalized finite difference method to seismic wave propagation in 2-D. *Journal of Computational and Applied Mathematics*, 236(12):3016–3025, 2011.

[2] E. Salete L. Gavete F. Ureña, J.J. Benito. Solving third and fourth order partial differential equations using GFDM, application to solve problems of plates. *International Journal of Computer Mathematics*, 89(3):366–376, 2012.

[3] L. Gavete F. Ureña, J.J. Benito. Influence several factors in the generalized finite difference method. *Applied Mathematical Modeling*, 25:10391053, 2001.

[4] L. Gavete R. Alvarez F. Ureña, J.J. Benito. An h-adaptive method in the generalized finite difference. *Comput. Methods Appl. Mech. Eng*, 192:735759, 2003.

[5] L. Gavete B. Alonso J.J. Benito, F. Ureña. Application of the generalized finite difference method to improve the approximated solution of pdes. *Computer Modelling in Engineering & Sciences*, 38:3958, 2009.

[6] Orkisz J. Liszka T. The finite difference method at arbitrary irregular grids and its application in applied mechanics. *Computer & Structures*, 11:8395, 1980.

[7] J. Orkisz. *Finite Difference Method (Part, III) in handbook of Computational Solid Mechanics*. M. Kleiber (Ed.) Springer-Verlag, Berlin 1998.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Windows for escaping particles in quartic galactic potentials

Juan F. Navarro[1]

[1] *Department of Applied Mathematics, University of Alicante*

emails: jf.navarro@ua.es

Abstract

We investigate the shape of the windows through which stars may escape from a galaxy modelled by a bi–symmetrical potential made up of a two–dimensional harmonic oscillator with quartic perturbing terms. It is well known that the escape from the potential well is governed by the unstable periodic orbits in the openings of the potential. The unstable and stable manifolds to these periodic orbits reveal the way test particles escape from the potential well. Our main objective is to compute accurately these manifolds to analyze the shapes and sizes of the windows of escape, founding that they consist of a "main window" and of a hierarchy of secondary spiral windows. This study is performed through the study of intersections of stable and unstable manifolds in the $x - \dot{x}$ Poincaré phase plane.

Key words: Galactic potentials, Periodic orbits, Escapes, Hamiltonian systems

1 Introduction

The phenomenon of escapes from a dynamical system, especially the escape of stars from stellar systems has been an active field of research during the last decades. In [2], Contopoulos explored escapes from dynamical systems representing the central part of perturbed galaxies, with Hamiltonian

$$\mathcal{H} = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \frac{1}{2}(x^2 + y^2) - \mu x^2 y^2 \,. \tag{1}$$

This system have quadruple symmetry and four openings to infinity. Every opening is bridged by an unstable periodic orbit called a Lyapunov orbit, which governs the escape to infinity from the potential well. The asymptotic curves of the Lyapunov periodic orbits make infinite rotations around some "limiting curves". The proportion of escaping orbits and the

 ISBN: 978-84-608-6082-2

direction of escape depend on the topology of the asymptotic surfaces of the Lyapunov periodic orbits. This is investigated in [3] by Contopoulos and Kauffman. The "basins" of escape toward different directions, and of the fast and the slow escapes for various values of the perturbation parameter (μ) are determined. In 1996, Siopis et al. [9] performed a numerical study of the escape properties of three two–dimensional, time–independent potentials possessing different symmetries. It was found, for all three cases, that (i) there is a rather abrupt transition in the behaviour of the late–time probability of escape, when the value of a coupling parameter, μ, exceeds a critical value, μ_2. For $\mu > \mu_2$, it was found that (ii) the escape probability manifests an initial convergence towards a nearly time–independent value, $p_0(\mu)$, which exhibits a simple scaling that may be universal. However, (iii) at later times the escape probability slowly decays to zero as a power–law function of time. Finally, it was found that (iv) in a statistical sense, orbits that escape from the system at late times tend to have short time Lyapunov exponents which are lower than for orbits that escape at early times. Navarro and Henrard [8] examined the shape of the windows through which test particles may escape in the simplified Hamiltonian system (1). Here, the authors analyzed the shapes and sizes of the windows of escape of stars from a simplified galactic model, founding that they consist of a "main window" and of a hierarchy of secondary windows. A very large part of the main window is actually made of "just passing through" stars and may not be very interesting for galactic studies. Hence the importance of the secondary windows, their intricate spiral structures and the fractality of the basin boundaries.

Later, in 2004, Contopoulos et al. [4] studied in detail the form of the asymptotic manifolds of a central unstable periodic orbit. The form of the escape regions and the infinite spirals of the asymptotic manifolds around the escape regions were given, as well as the computation of the escape rate for different values of the energy and the percentage of orbits that escape after a finite number of iterations. The problem of the escape in galactic potentials has been recently revisited by Zotos ([10], [11]), who has investigated the structure of the phase space of the dynamical system described by (1).

The aim of this work is to better understand the properties of the escape of orbits in a simple local galactic Hamiltonian describing the motion near the centre of a elliptical galaxy. In the present work, we study the properties of the escape in the galactic system described by the potential

$$W(x,y) = \frac{1}{2}(\omega_1^2 x^2 + \omega_2^2 y^2) - \mu\left[\beta(x^4+y^4) + 2\alpha x^2 y^2\right] , \tag{2}$$

where ω_1, ω_2 are the unperturbed frequencies of oscillation along the x and y axis respectively, $\mu > 0$ is the perturbation strength, and α and β are parameters. We shall study the case $\omega_1, \omega_2 = \omega = 1$, that is, the $1:1$ resonance case. The Hamiltonian to the potential (2) is [1]

$$\mathcal{H} = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \frac{1}{2}(x^2+y^2) - \mu\left[\beta(x^4+y^4) + 2\alpha x^2 y^2\right] . \tag{3}$$

We will compute the asymptotic manifolds to the Lyapunov periodic orbits in the openings of the potential in order to show that the mechanism described in [8] is valid also in this case. As in the simplified model described there, we show that the intricate spiral structures of the secondary windows governs the rate of escape: there are "first order", infinitely winding spirals, but also "second order" spirals composed themselves of an infinity of layers, "third order" spirals formed by an infinity of second order spirals, and so on.

Hence, in order to investigate the size, shape and properties of the regions of phase space leading to escape, it is necessary to understand the geometry of the stable manifolds to the "guardian" periodic orbits. We will do so by investigating the intersections of the stable and unstable manifolds with a surface of section. This has been the strategy of Contopoulos and coworkers ([2], [3]) and Navarro and Henrard ([8]).

Here, we revisit this method to show the relation between the infinite spiraling and the escape. Furthermore, we show that, again, there are first order spirals, but there are also second order spirals which are formed of an infinite number of spirals, embedded in each other like Russian dolls, and then third order spirals, formed by an infinite number of second order spirals, and so on.

In our analysis, we will take as surface of section the plane $y = 0$, and use a sixth order expansion of the stable and unstable manifolds in order to obtain precise initial conditions far enough from the unstable "guardian orbit" to start a safe computation of these manifolds.

2 Symmetries of the Problem

There are three different cases of the $x - \dot{x}$ phase portrait in the Hamiltonian (1): (a) $\alpha > \beta > \alpha/3$ ($\alpha > 0$, $\beta > 0$). (b) $\beta > \alpha$ ($\alpha > 0$, $\beta > 0$). (c) $\alpha > 3\beta$ ($\alpha > 0$, $\beta > 0$), or $\alpha < 0$, $\beta > 0$. In all these three situations, there is, for a fixed value of the energy, a critical value of the energy (h_c) such that, for larger values of h, the potential well opens up to infinity and test particles may escape. When $\alpha > \beta > \alpha/3$ ($\alpha > 0$, $\beta > 0$), this critical value is given by

$$h_c = \frac{1}{8\mu(\alpha + \beta)} \, .$$

In our analysis, we will consider the following values of the parameters of the system: $\mu = 2.64$, $\alpha = 1.2$ and $\beta = 0.8$. The critical value of the energy associated to these values of the parameters is $h_c = 0.0236742$. For each larger value of h, there is an unstable periodic orbit across the opening, bouncing back and forth between the two "walls" of the pass (see Figure 1).

It has been long recognized ([2], [8]) that these unstable periodic orbits are, in some sense, the guardians of the pass: orbits going through the pass are sheperded by the stable and unstable manifolds of the periodic orbit. This can be explained in the following way: a continuous variation of initial conditions may lead from an orbit which "bounces back" in

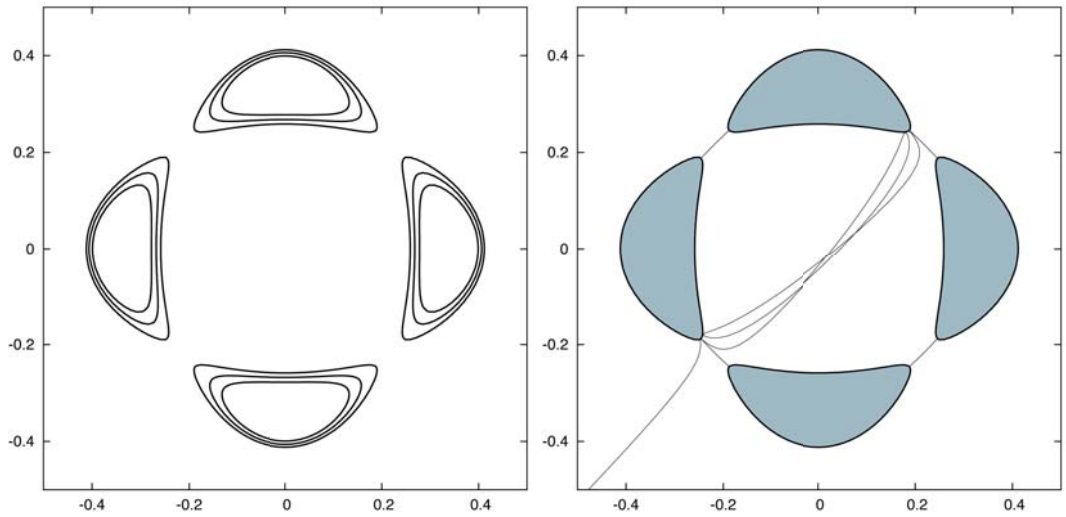


Figure 1: Curves of zero velocity for $\mu = 2.64$, $\beta = 0.8$, $\alpha = 1.2$ and different values of the energy larger than the critical value: $h_1 = 0.024$, $h_2 = 0.025$ and $h_3 = 0.026$ (left panel). As the energy of the system grows, the windows of the potential well become wider. In the right panel, we show an orbit belonging to the stable manifold to the periodic orbit "guarding" the upper-right opening in the potential well for $\mu = 2.64$, $\beta = 0.8$, $\alpha = 1.2$ and $h = 0.024$. The guardian orbit is the almost straight line barring the opening. The orbit belonging to the stable manifold is computed backward in time starting from initial conditions computed from a sixth order approximation.

the pass and returns inside the well to an orbit which passes through, a very non–continuous behavior. The only way this can happen is that somewhere between these two orbits there is a third one which stays in the pass for an infinite amount of time. Indeed, the general solution $X(x(0), t)$ is a continuous function of the initial condition $x(0)$ for any *finite* time t.

Due to the symmetries of the potential, the well opens up, for values of the energy larger than the critical value h_c, at four places along the diagonals of the configuration space ($x = \pm y$) (left panel of Figure 1). The phase space of Hamiltonian (1) is symmetric with respect to each of the two coordinates (x, y) (and their velocities) and with respect to time (and velocities) reversal, i.e. if $(x(t), y(t), \dot{x}(t), \dot{y}(t))$ is an orbit, then the following are

also orbits

$$\begin{array}{lrrrr}
\text{(a)} & x(t), & y(t), & \dot{x}(t) & ,\dot{y}(t)\,, \\
\text{(b)} & -x(t), & y(t), & -\dot{x}(t) & ,\dot{y}(t)\,, \\
\text{(c)} & x(t), & -y(t), & \dot{x}(t), & -\dot{y}(t)\,, \\
\text{(d)} & -x(t), & -y(t), & -\dot{x}(t), & -\dot{y}(t)\,, \\
\text{(e)} & x(-t), & y(-t), & -\dot{x}(-t), & -\dot{y}(-t)\,, \\
\text{(f)} & -x(-t), & y(-t), & \dot{x}(-t), & -\dot{y}(-t)\,, \\
\text{(g)} & x(-t), & -y(-t), & -\dot{x}(-t), & \dot{y}(-t)\,, \\
\text{(h)} & -x(-t), & -y(-t), & \dot{x}(-t), & \dot{y}(-t)\,.
\end{array} \tag{4}$$

Hence it is enough to compute the sections of the stable manifold to one periodic orbit (let us take as (a) the stable manifold to the periodic orbit in the upper–right corner) in order to know by symmetry the sections of the stable and unstable manifolds of the four guardian periodic orbits. This is shown in Figure 2.

3 Computation of the Stable and Unstable Manifolds

Due to the instability of the periodic orbit, we have determined one of them for a value of the energy slighty larger than the critical value following a geometrical approach, and then we have continued the family of periodic orbits taking the energy as parameter. The periodic orbit analyzed corresponds to $x_0 = y_0 = 0.2165477494383151634$, $\dot{x}_0 = -\dot{y}_0 = -0.0364410361567080940$, and $T/2 = 5.6343161357767205244$, with energy $h = 0.0249$.

As mentioned in the introduction, the stable manifolds to the "guardian" unstable periodic orbits form the boundaries of the escape windows we wish to analyze. For the computation of the initial part of these stable manifolds (and of the unstable manifolds as well), we shall follow the scheme proposed by Deprit and Henrard [6]. The stable manifold to a periodic orbit $(x^*(t), y^*(t))$ of period T is represented by series

$$\begin{aligned}
x(t,\epsilon) &= x^*(t) + \epsilon u(t) \quad , \quad u(t) = \sum_{j\geq 1} \epsilon^{j-1} e^{-jat} x_j(t)\,, \\
y(t,\epsilon) &= y^*(t) + \epsilon v(t) \quad , \quad v(t) = \sum_{j\geq 1} \epsilon^{j-1} e^{-jat} y_j(t),
\end{aligned} \tag{5}$$

where the coefficients $x_j(t)$ and $y_j(t)$ are periodic with period T and a is the (positive) characteristic exponent of the periodic orbit. The unstable manifold is represented by similar series with the exponent $-a$ replaced by $+a$. When ϵ is equal to zero, we recover the periodic orbit. Substituting the series (5) in the differential equations derived from the Hamiltonian (3), we obtain

$$\ddot{u} \;=\; W^*_{xx} u + W^*_{xy} v + \sum_{i\geq 1} \epsilon^{i-1} \sum_{j\geq 0} c_i^j \left(\frac{\partial^i W}{\partial x^j \partial y^{i-j}} \right)^* u^j v^i\,,$$

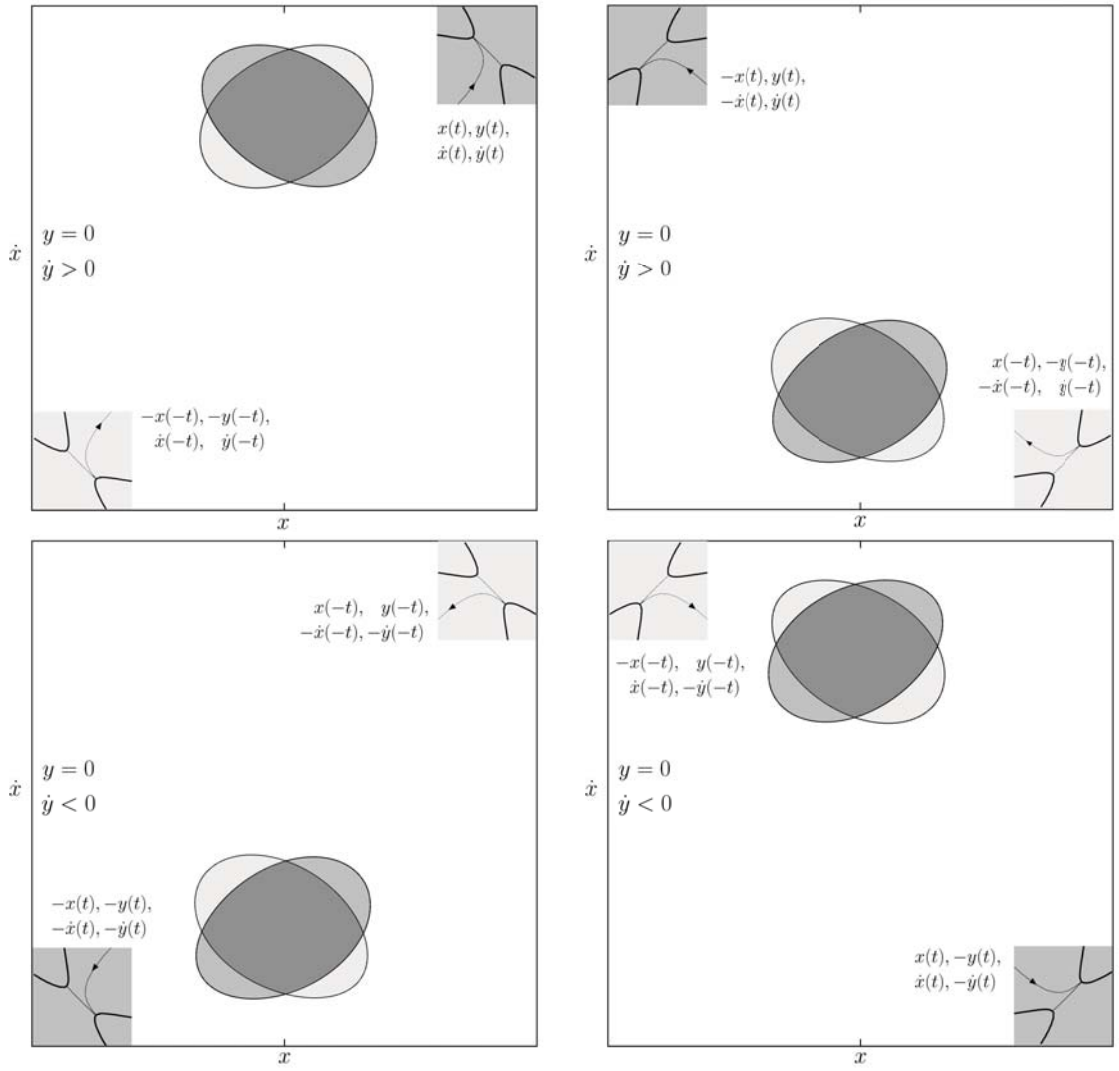


Figure 2: The first intersections of the two surfaces of section ($y = 0$, $\dot{y} > 0$) and ($y = 0$, $\dot{y} < 0$) with the symmetrics (enumerated in equation 4) of the (internal part of the) stable manifold. The areas in medium grey correspond to orbits leaving the potential well. Areas in light grey correspond to orbits coming from "outside" and which, as a consequence, do not have any antecedents on the surfaces of section. Of course areas with dark grey correspond to orbits "just passing through" the potential well.

$$\ddot{v} \;=\; W^*_{yx}u + W^*_{yy}v + \sum_{i\geq 1}\epsilon^{i-1}\sum_{j\geq 0} c_i^j \left(\frac{\partial^{i+j+1}W}{\partial x^j \partial y^{i-j}}\right)^* u^j v^i\,, \qquad (6)$$

where we have used the notation f^* for the function f evaluated along the periodic solution $[f^*(t) = f(x^*(t), y^*(t))]$. Expanding $u(t)$ and $v(t)$ in powers of ϵ, the equation (6) lead to a recursive set of linear equations for the coefficients $x_k(t)$ and $y_k(t)$.

4 The Boundaries of the Escape Windows

We have integrated numerically and backward (using Lie series of order 20 [7]) some 1 000 000 orbits belonging to the stable manifold to the first quadrant guardian, until they cross the hyperplane $y = 0$ (with $\dot{y} > 0$). The initial conditions were computed by keeping the value of $\epsilon \exp(at)$ fixed at 0.0001 and choosing 1 000 000 equidistant values of the initial time t.

In Figure 3 (left panel), we show the first section of the stable manifold to the first quadrant guardian periodic orbit for $h = 0.0249$. By symmetry we can also plot the trace of the unstable manifold to the third quadrant guardian (right panel). We observe that the two "rings" intersect. Orbits starting inside both rings are coming from the lower–left infinity and are going to the upper–right infinity. They just pass through the center of the galaxy. On the other hand, orbits starting from one of the two crescents shown in the right panel of Figure 3 originate inside the galaxy (and of course escape in the future by the upper–right window), and we can consider their previous intersections with the surface of section.

The second intersection, shown in Figure 4 (left panel), is composed of two tongues which spiral around the stable manifold to the third quadrant guardian orbit. The spirals are infinite, but of course we have computed (and shown in Figure 4) only a part of them. We show also in Figure 4 the ring created by the unstable manifold to the first quadrant orbit (Figure 2). Orbits starting inside this ring, and inside one of the two tongues, come from infinity by the upper–right window and leave the galaxy by the same window after two crossings of the axis $y = 0$. Orbits inside the tongues but outside the light blue area of Figure 4 (right panel) have a previous intersection with the surface of section.

Following these latter orbits backward, we compute the third section (see figure 5). It is composed of two "simple" tongues. The tongues are "Russian dolls", composed of "subtongues" embedded inside each other. Only a few of these subtongues have been computed, but there are an infinity of those subtongues. All these tongues are infinitely winding around the stable manifold to the first quadrant guardian (already shown in Figure 3). Orbits starting inside this later manifold (in the area shown in dark blue grey) are those which escape directly by the first quadrant window.

The fate of the orbits starting inside the complex tongues alternates according to the parity of the number of subtongues in which they are embedded. Those starting from the

 ISBN: 978-84-608-6082-2

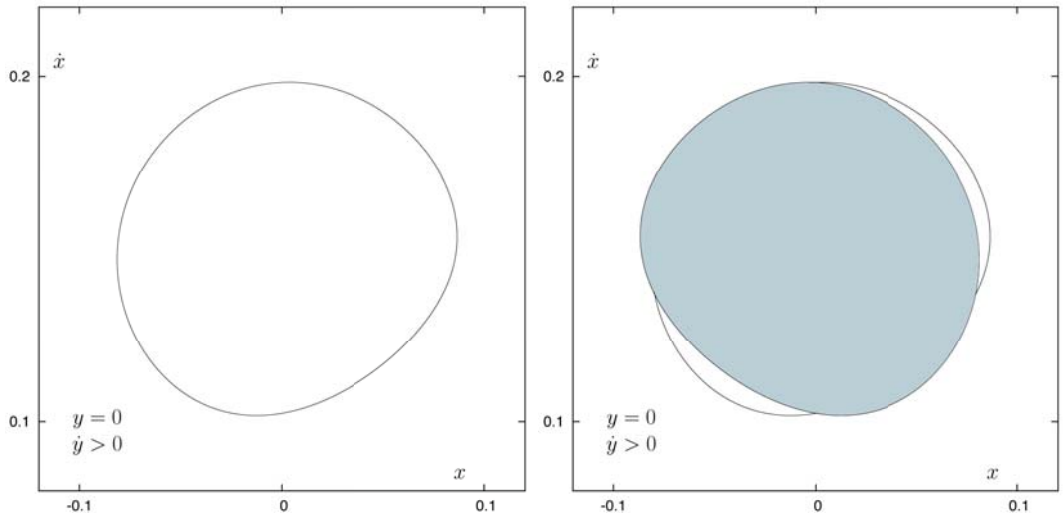


Figure 3: First section ($y = 0$, $\dot{y} > 0$) of the stable manifold to the first quadrant guardian periodic orbit (on the left panel). On the right panel we have also reproduced (in light blue) the section of the unstable manifold to the third quadrant periodic orbit. The area inside both sections corresponds to orbits coming from infinity from lower left, passing through the potential well once, and dissppearing to infinity to the upper right. The two remaining crescents (in white) will have antecedents on the surface of section (Figure 4).

 ISBN: 978-84-608-6082-2

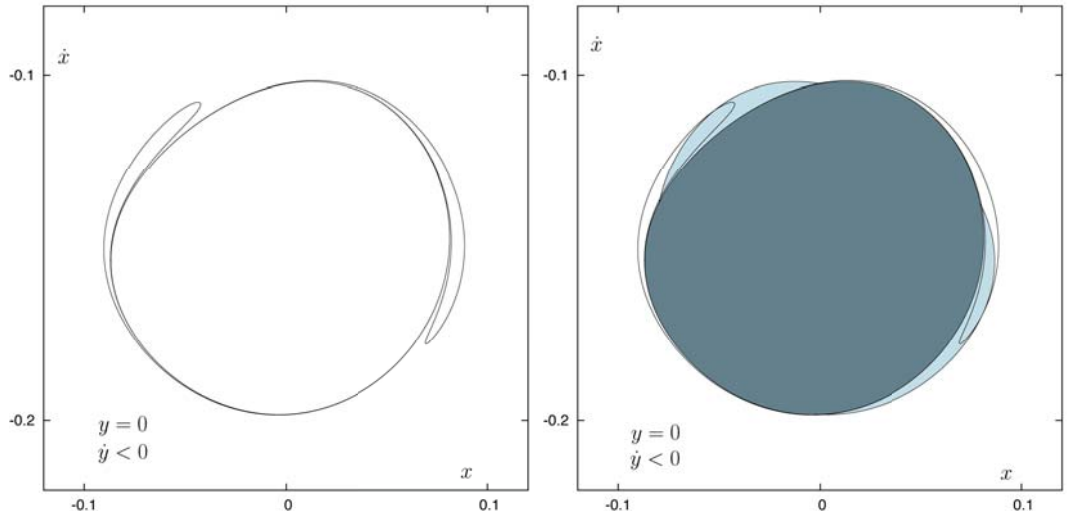


Figure 4: Second section ($y = 0$, $\dot{y} < 0$) of the stable manifold to the upper–right guardian periodic orbit (left panel). The section is composed of two tongues, images of the two crescents of the right panel of Figure 3, winding around the stable manifold to the third quadrant periodic orbit (shown in dark blue). Also shown (in light blue) is the section of the unstable manifold to the upper–right guardian periodic orbit.

 ISBN: 978-84-608-6082-2

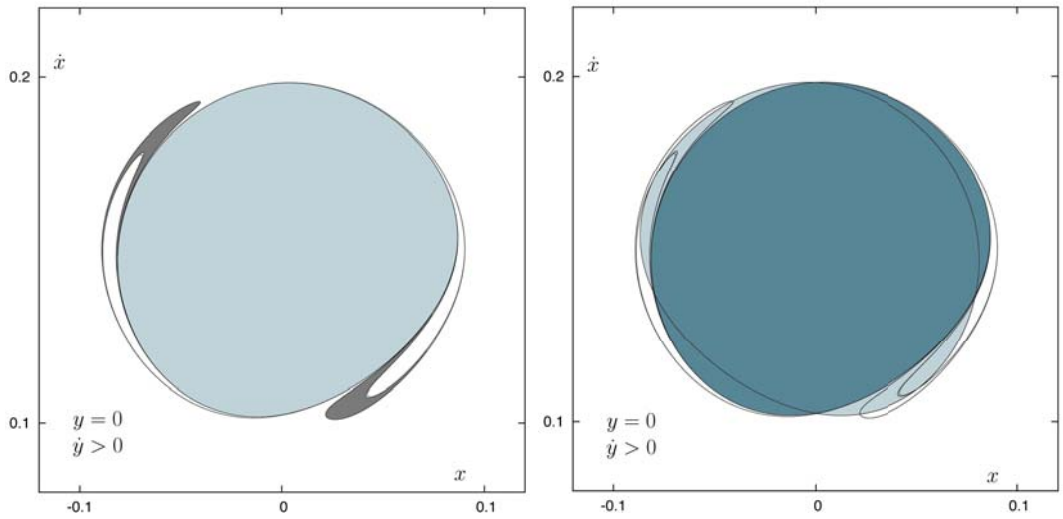


Figure 5: Third section (bt $y = 0$, $\dot{y} > 0$) of the stable manifold to the first quadrant periodic orbit. The section is composed of two complex tongues, "russian dolls" composed of an infinity of "subtongues" embedded inide each other. All these tongues wind around the stable manifold to the first quadrant (in light blue in left panel). In the right panel, we also represent in dark blue the unstable manifold to the third quadrant periodic orbit.

area shown in grey in Figure 5 are mapped on the pieces of the tongues of section 2 which are not inside the light blue grey area (right panel of Figure 4) and thus escape by the upper–right window after two crossing of the surface of section.

We also reproduce on Figure 5 the trace of the unstable manifold to the third quadrant periodic orbit (shown in light blue grey in the right panel of Figure 3). Orbits inside it and inside the tongues come from infinity by the lower–left window; the others have antecedents on the surface of section and we could compute them backward to a fourth intersection with the surface of section. The result of this intersection is shown in Figure 6. As we see, the scheme is the same as before.

5 Conclusions

In this paper, we study the properties of the escape of orbits in a simple Hamiltonian model describing the motion near the centre of an elliptical galaxy. For certain values of the energy, the potential well opens up to infinity and test particles may escape. Every opening is bridged by an unstable periodic orbit called a Lyapunov orbit, which governs the escape to infinity from the potential well. The sets of escaping orbits are limited by the stable

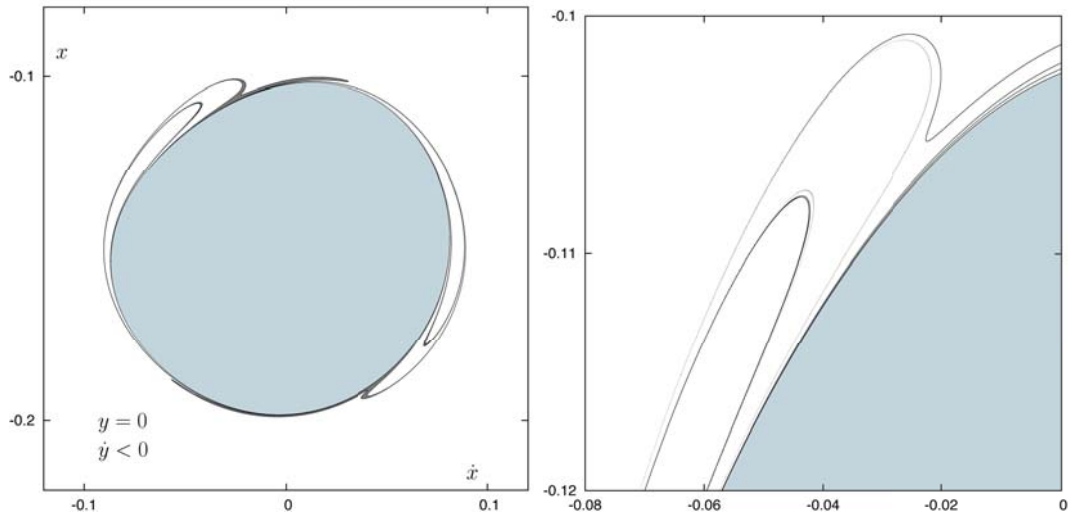


Figure 6: Fourth intersection (by $y = 0$, $\dot{y} < 0$) of the stable manifold to the first quadrant periodic orbit. As in Figure 5, the section is composed by complex tongues composed of an infinity of subtongues embedded inside each other. All these tongues wind around the stable manifold to the third quadrant periodic orbit.

and unstable manifolds to these periodic orbits. In [8], Navarro and Henrard computed the intersections of these manifolds with a surface of section in order to clarify the shapes and sizes of the windows of escape of stars. We found that in the system given by Hamiltonian (1), these windows consist of a "main window" and of a hierarchy of secondary windows. A very large part of the main window is actually made of "just passing through" stars and may not be very interesting for galactic studies. Hence the importance of the secondary windows, their intricate spiral structures and the fractality of the basin boundaries.

Here, we have analyzed the same problem in a more complex Hamiltonian, given by (3). We have also investigated the intersections of the stable and unstable manifolds to the guardian periodic orbits with a surface of section, just to find the same mechanism governing the escape from the potential well. As in Hamiltonian (1), we have shown that the intricate spiral structures of the secondary windows governs the rate of escape: there are "first-order", infinitely winding spirals, but also "second order" spirals composed themselves of an infinity of layers, "third order" spirals formed by an infinity of second order spirals, and so on.

However, we have found some differences in the number of "main" tongues in the third intersection of the stable manifold to the upper–right guardian periodic orbit with the surface of section $y = 0$, $\dot{y} > 0$. This is a consequence of the geometric shape of the tongues in the second section ($y = 0$, $\dot{y} < 0$) of the stable manifold to the upper–right guardian

periodic orbit (see Figure 4, left panel), and the way this curve intersects with the unstable manifold to the upper–right guardian periodic orbit. In the case of Hamiltonian (1), the intersections originate four tongues, two simple and two complex. Here, there are only two complex tongues, composed of "subtongues" embedded inside each other. In spite of the fact that these geometric differences between both problems exist, we have found the same mechanism to explain the escape of stars from the potential well.

References

[1] N. D. Caranicolas, Ch. L. Vozikis, *Order and chaos in galactic maps*, Astron. Astrophys. **349** (1999) 70.

[2] G. Contopoulos, *Asymptotic curves and escapes in Hamiltonian systems*, Astron. Astrophys. **231 (1)** (1990) 41–45.

[3] G. Contopoulos, D. Kaufmann, *Types of escapes in a simple Hamiltonian system*, Astron. Astrophys. **253 (2)** (1992) 379–388.

[4] G. Contopoulos, K. Efstathiou, *Escapes and Recurrence in a Simple Hamiltonian System*, Celest. Mech. Dyn. Astron. **88 (2)** (2004) 163–183.

[5] A. Deprit, J. Henrard, *A manifold of periodic orbits*, Adv. in Astron. and Astrophys. **6** (1968) 1–124.

[6] A. Deprit, J. Henrard, *Construction of orbits asymptotic to a periodic orbit*, Astron. J. **74** (1969) 308–316.

[7] A. Deprit, J. F. Price, J.F., *Numerical integration by recurrent power series*, Astron. Astrophys. **1** (1969) 427.

[8] J. F. Navarro, J. Henrard, *Spiral windows for escaping stars*, Astron. Astrophys. **369** (2001), 1112–1121.

[9] C. Siopsis, H. E. Kandrup, G. Contopoulos and R. Dvorak, *Universal properties of escape in dynamical systems*, Celest. Mech. Dyn. Astron. **65(1-2)** (1996) 57–68.

[10] E. E. Zotos, *Trapped and escaping orbits in an axially symmetric galactic–type potential*, PASA **29** (2012) 161–173.

[11] E. E. Zotos, *Escape dynamics in a Hamiltonian system with four exit channels*, Nonlinear Studies **22 (3)** (2015) 1–20.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Accelerating Microrheology models on HPC architectures

Gloria Ortega[1], Antonio M. Puertas[2] and Ester M. Garzón[1]

[1] *Group of Supercomputation-Algorithms, Dpt. of Informatics, ceiA3, Univ. of Almería, 04120, Almería, Spain*

[2] *Group of Complex Fluids Physics, Dpt. of Applied Physics, Univ. of Almería, 04120, Almería, Spain*

emails: gloriaortega@ual.es, apuertas@ual.es, gmartin@ual.es

Abstract

Complex fluids are characterized with both solid and fluid proprieties by their elasticity and viscosity, or rheological properties. The flow of complex fluids is a focus of interest in a very wide range of applications in biophysics and soft matter, such as problems involving live cells, processing of plastic, glass, paints, foods, oil recovery and so on. The development of these applications needs the estimation of the rheological properties of soft materials. Recently, microrheology has been developed as an accurate technique to obtain rheological properties in soft matter from the microscopic motion of colloidal tracers used as probes, either freely diffusing in the host medium (passive), or subjected to external forces (active). A drawback for these techniques is their hight computational cost. Therefore, the use of high performance computing is mandatory to develop the microrheology models. In this work, the parallelisms involved in the microrheology computation and the heterogeneous platforms which optimize the performance of these models are analysed.

Key words: microrheology, heterogeneous computation, high performance applications

1 Introduction

Rheology is the study of the flow properties of matter, particularly of soft matter, which shows complex behaviour, identified by non-linear curves in the stress vs. velocity of strain; typical examples are paints, corn starch, polymer melts and so on. It is therefore an interdisciplinary subject, spread between different communities including chemical engineers,

physicists, material scientists and chemists [1]. Rheological models try to provide (macroscopic) constitutive equations for a particular material based on first principles at the microscopic level. Microrheology, on the other hand, was proposed almost twenty years ago as an alternative technique to obtain rheological properties in expensive, or difficult to obtain samples. In microrheology, one uses the microscopic motion of colloidal tracers used as probes, either freely diffusing in the host medium, or subjected to external forces [6], but the theoretical analysis is rather involved.

A relevant drawback of the Microrheology models is their high computational costs [1]. From the computational point of view, these models are defined as a hierarchy of tasks with high computational cost which can be executed in parallel. Therefore, the use High Performance Computing (HPC) can help scientists and engineers to develop and apply Microrheology models. It is due to the HPC can strongly reduce the execution time of them.

Nowadays, HPC architectures are characterized by the heterogeneity of their resources. The modern supercomputers consist of clusters of multi-core nodes which include accelerator devices such as GPUs [1]. Current stand-alone computers present tremendous power, thanks to technological and architectural advances and they could be considered as desktop supercomputers if their heterogeneous resources could be appropriately exploited [4]. These architectures, based on multi-core processors, belong to the classical category of shared-memory computers in HPC. There are standard paradigms and programming tools that allow programmers to fully take advantage of their power easily. Usually, they also include GPUs as accelerator resources to take advantage of the several kinds of parallelism in applications. GPUs offer massive parallelism and provide outstanding performance-to-cost ratio for scientific computing.

Therefore, modern architectures provide the resources which are required by the Microrheology models. In this work, a particular active Microrheology model is analysed and taken as a prototype in this field. Colloidal systems are considered as hard spheres systems whose rheological are computed by the study of the dynamic of a tracer particle. Several parallelism levels are identified and related to the computational resources which allow their parallel execution. The bottom level parallelism which can be managed on the GPUs platforms is our main focus of interest. So, a GPU version of the procedure to compute the tracer trajectory is evaluated.

2 Description of the problem

The study of the rheological properties of soft matter at the microscopic scale using colloidal tracers, is termed microrheology. Here, we tackle the problem of microrheology in colloidal

[1] http://www.top500.org/lists/2015/11/

 ISBN: 978-84-608-6082-2

systems of hard spheres by means of particle simulations, and study the applicability of GPU programming in the acceleration of the obtention of the tracer trajectory.

In the simulation of Brownian systems, a friction force with the solvent and a random force must be considered in addition to the direct interaction forces. The friction force is proportional to the particle-particle velocity, and the random force is linked to the friction coefficient via the fluctuation dissipation theorem [2]. The final equation for particle j reads,

$$m\frac{d^2\vec{r}_j}{dt^2} = \sum_i \vec{F}_{ij} - \gamma_0\frac{d\vec{r}_j}{dt} + \vec{\eta}_j(t) \quad \left(+\vec{F}_{\text{ext}} \right) \tag{1}$$

The last term is a constant external force that is applied only to the tracer particle (all particles undergo Brownian motion, but only the tracer is pulled). The simulations are run in a cubic box, with N particles and periodic boundary conditions. All bath particles have the same mass, $m = 1$, and radius, $a = 1$, and the thermal energy is $k_{\text{B}}T = 1$. The tracer has mass $m_t = 1$ and radius $a_t > a$. The density of the bath is given by the fraction of the volume occupied by the particles, $\phi = 4/3\pi a^3 n$, where n is the number density; in our system the volume fraction is fixed to $\phi = 0.50$. The solvent friction coefficient of particle i is set to $\gamma_0 = 5a_i\sqrt{mk_{\text{B}}T}/d$. Time is measured in units of the bath Newtonian microscopic time $a\sqrt{m/k_{\text{B}}T}$. The equations of motion are integrated with a time step of $0.0005a\sqrt{m/k_{\text{B}}T}$ using an extension of the velocity Verlet algorithm to include random forces [5], integrating the friction forces analytically.

Particle-particle interactions are given by an inverse power potential, with a large exponent; $V(r_{ij}) = k_{\text{B}}T(r/d_{ij})^{-36}$, where d_{ij} is the center to center distance ($d_{ij} = (a_i + a_j)$, where a_i is the diameter of particle i. Due to the short range of this interaction potential, its cut-off radius is set to $a_i + a_j + a$, what allows the use of typical procedures to speed up the calculation, such as cell or Verlet lists. Despite this, the calculation of the forces is the most time-consuming part of the code.

In our simulations of microrheology, the system with the (large) tracer particle is equilibrated, and at time $t = 0$ the tracer is pulled with a constant force through the system. The main output of the simulation is the trajectory of the tracer, that yields the effective friction coefficient from its average velocity, using the relation $\vec{F}_{\text{ext}} = \gamma_{\text{eff}}\langle\vec{v}\rangle$, valid for the stationary regime. The tracer is allowed to travel through the simulation box more than once (obeying the periodic boundary conditions), as we could not identify any different behaviour between the first and consecutive passages. Fig. 1 shows several individual trajectories, and the inset presents a snapshot of the system.

Because the tracer is much larger than the bath particles, finite size effects are possible. The periodic boundary conditions in fact simulate the dynamics of a cubic array of tracer particles immersed in a Brownian bath. Theoretical analysis within the (continuous model) Navier-Stokes equation predicts a linear dependence of the inverse friction coefficient with the inverse system size, L, for vanishing forces [3]:

 ISBN: 978-84-608-6082-2

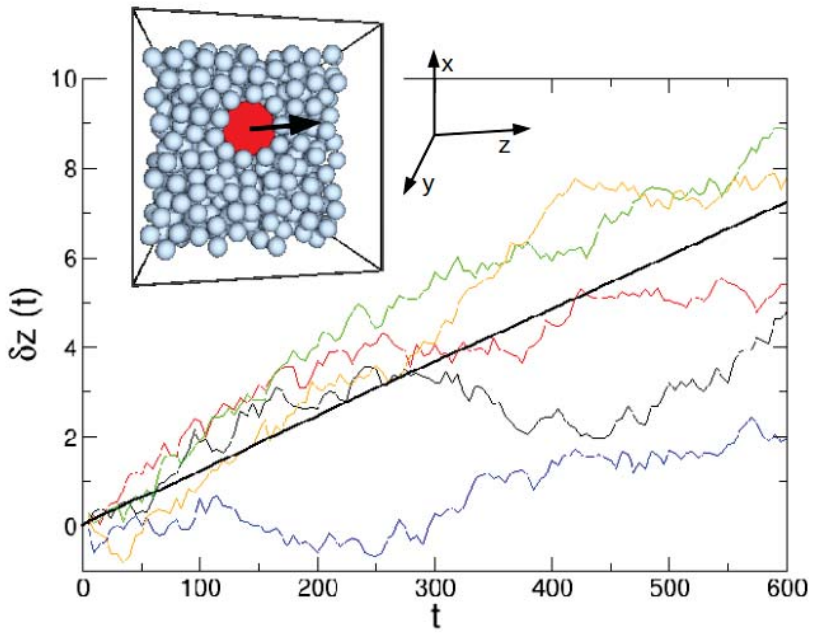


Figure 1: Five trajectories of the tracer (thin lines) and average trajectory (thick line). $\delta z(t)$ axis represents the tracer advance in the $\vec{F}_{\text{ext}}$ direction at the instant t. Inset: Snapshot of the system with the tracer marked in red, and all particles in front of it removed to allow observing the tracer. The arrow indicates the external force applied only to the tracer.

 ISBN: 978-84-608-6082-2

$$\frac{1}{\gamma_{\text{eff}}} - \frac{1}{\gamma_\infty} \propto \frac{1}{L} \tag{2}$$

where γ_∞ is the effective friction coefficient for an infinitely large system, which is the quantity we are seeking. The procedure to obtain γ_∞ from simulations is thus to run simulations with different system sizes and extrapolate $1/\gamma_{\text{eff}}$ linearly for $1/L \to 0$. This extrapolation is tricky as small errors in the determination of γ_{eff} for finite L imply large errors in γ_∞ given the long distance between the data and the $1/L = 0$ axis. It is therefore desirable to run simulations with L as large as possible, in a reasonable time, to reduce this error. Notice that $N = 3\phi L^3/(4\pi a^3)$.

3 Computational analysis and parallel implementation

The goal of the model described in Section 2 is the computation of the macroscopic propriety of the colloidal system, γ_{eff}. So, to obtain γ_{eff} the following nested iterative procedures in bottom-up order have to be computed:

- A tracer trajectory in the $\vec{F}_{\text{ext}}$ direction, referred as $TT_i(N)$ for a specific number of particles and its corresponding system size, L.
- The average of $TT(N)$ according to random decisions (i) to advance in the $\vec{F}_{\text{ext}}$ direction, for a specific value of N.
- The average of $TT(N)$ for several values of N to extrapolate γ_{eff} according to Equation 2.

Therefore, there are three parallelism levels in the model. In this work, our focus is to accelerate the computation of every tracer trajectory, at bottom level. Then, a time stepping procedure is applied to compute the location of the tracer particle after a differential time interval. At every time step, two main tasks are completed: (1) identifying of the neighbors of N particles; and (2) evaluating of the locations and velocities of the N particles according to the particle-particle interactions among the neighbor particles defined by Equation 1. The complexities of both tasks are $O(N^2)$ and $O(N)$, respectively. So, the first task consumes most of the time. However, it is only computed when a relevant movement of particles is detected in the previous time step.

As above mentioned, to achieve accurate values of γ_{eff}, it is mandatory the study of simulations for large L values, that is, large particles number in the system (N). For large values of N, it is worth to accelerate both tasks on the GPU platforms.

Therefore, Microrheology problems with several sizes have been executed and evaluated on a platform with a GPU. Obtained results in terms of performance have shown the advantages of the GPU computing to accelerate this kind of problems when N is large enough.

 ISBN: 978-84-608-6082-2

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science throughout projects TIN2012-37483, TIN15-66680, FIS-2015-69022-P and CAPAP-H5 network TIN2014-53522, by J. Andalucía through projects P12-TIC-301 and P11-TIC7176, and by the European Regional Development Fund (ERDF).

References

[1] P. Cicuta and A.M. Donald. Microrheology: a review of the method and applications. *Soft Matter*, 3:1449–1455, 2007.

[2] J.K.G. Dhont. *An Introduction to Dynamics of Colloids*. Studies in Interface Science. Elsevier Science, 1996.

[3] H. Hasimoto. On the periodic fundamental solutions of the Stokes equations and their application to viscous flow past a cubic array of spheres. *Journal of Fluid Mechanics*, 5:317–328, 1959.

[4] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, 2011.

[5] W. Paul and D. Y. Yoon. Stochastic phase space dynamics with constraints for molecular systems. *Phys. Rev. E*, 52:2076–2083, Aug 1995.

[6] A.M. Puertas and T. Voigtmann. Microrheology of colloidal systems. *Journal of Physics: Condensed Matter*, 26(24):243101, 2014.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Argon Melting under High Pressure

Elke Pahl[1], Jonas Wiebke[1], Florian Senn[1] and Peter Schwerdtfeger[2]

[1] *Centre for Theoretical Chemistry and Physics, Institute for Natural and Mathematical Sciences (INMS), Massey University Albany, Private Bag 102904, Auckland 0745, New Zealand*

[2] *Centre for Theoretical Chemistry and Physics, The New Zealand Institute for Advanced Study (NZIAS), Massey University Albany, Private Bag 102904, Auckland 0745, New Zealand*

emails: e.pahl@massey.ac.nz, , , p.a.schwerdtfeger@massey.ac.nz

Abstract

Our goal is to gain a detailed understanding of melting processes occurring under high pressure. As a first model system we have chosen Argon as a (under ambient condition) very weakly bound system. I will give an overview about our approach while presenting results for the melting of bulk Ar as well as finite clusters of Ar atoms for pressures of up to 100 GPa.

At the heart of all studies is an accurate, while computationally efficient description of the interaction potential for systems of N atoms [1]. For rare gases a standard many-body expansion of the interactions works well, with the two-body data fitted to an extended Lennard-Jones form to guarantee for computational efficiency [2] . Summing over all two- and three body contributions (derived from highly accurate ab initio data) yields converged results. The dynamics is then performed on the accurate potential surfaces via classical Monte Carlo (MC) simulations. The parallel-tempering MC version allows for a simultaneous propagation at several temperatures and for information exchange between the runs at different temperatures by configuration swaps. This is a very efficient way to get the needed thermodynamics quantities, here the inner energy/enthalpy and heat capacities [3, 4].

For the pressure studies our MC melting approach was extended to an isobaric, isothermal ensemble. For an isobaric sampling, in addition to randomised atom moves one has to allow for volume moves, allowing the volume to adapt for the given pressures and temperatures. These volume moves are straightforward for the bulk which is modelled as an extended, infinite system via periodic boundary conditions: We can simply

 ISBN: 978-84-608-6082-2

vary the volume of the chosen simulation cell. For finite number clusters, however, the volume definition itself is not straightforward, and several possible volume definitions were tested against each other for clusters of several hundreds of atoms. These cluster sizes were shown previously to already allow for an extrapolation to the bulk limit in the ambient pressure regime [3, 4].

Results for the melting of solid Ar under high pressures of 1-100 GPa are found to be very accurate and probably exceeding experimental accuracy [5]. In contrast, for Ar clusters the need to include the environment as the medium to convey the pressure is demonstrated. Different volume definitions for the 'bare', environment-free clusters are working well for ambient and slightly elevated pressures up to the MPa regime, but start to fail spectacularly for pressures in the high pressure region. Possible ways forward are discussed.

Key words: Melting, Parallel-tempering Monte Carlo, High Pressure, Rare Gases

Acknowledgements

This work has been supported by the Marsden fund administered by the Royal Society of New Zealand.

References

[1] E. Pahl, F. Calvo, and P. Schwerdtfeger, *The Importance of Accurate Interaction Potentials in the Melting of Argon Nanoclusters*, Int. J. Quant. Chem. 109 **9** (2009) 1812–1819.

[2] P. Schwerdtfeger, N. Gaston, R. P. Krawczyk, R. Tonner, and G. E. Moyano, *Extension of the Lennard-Jones potential: Theoretical investigations into rare-gas clusters and crystal lattices of He, Ne, Ar and Kr using many-body interaction expansions*, Phys. Rev. B **73** (2006) 064112–1–19.

[3] E. Pahl, F. Calvo, L. Koči, and P. Schwerdtfeger, *Accurate Melting Temperatures for Neon and Argon from Ab Initio Monte Carlo Simulations*, Angew. Chem. Int. Ed. **47** (2008) 8207–8210.

[4] F. Senn, J. Wiebke, O. Schumann, S. Gohr, P. Schwerdtfeger, and E. Pahl, *Melting of non-magic argon clusters and extrapolation to the bulk limit*, J. Chem. Phys. **140** (2014) 044325–1–5.

[5] J. Wiebke, E. Pahl, and P. Schwerdtfeger, *Melting at high pressure: Can first-principles computational chemistry challenge diamond-anvil cell experiments?*, Angew. Chem. Int. Ed. **52** (2013) 13202–13205.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Isolated inhomogeneities of arbitrary shape with polynomial fields prescribed at infinity. The potential problem in two dimensions.

William J. Parnell[1] **and Carmen Calvo-Jurado**[2]

[1] *School of Mathematics, Oxford Road, Manchester, M13 9PL, University of Manchester*

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

emails: William.Parnell@manchester.ac.uk, ccalvo@unex.es

Abstract

The so called Eshelby inhomogeneity or inclusion problem constitutes a crucial tool in the theory of composites and micromechanics [1]. Eshelby showed that the strain and stress interior to an isolated elliptical (2D) or ellipsoidal (3D) inhomogeneity embedded in an infinite homogeneous medium Ω subjected to uniform strain or traction in the far field, will also be uniform. This property, also called the *Eshelby uniformity property*, has been studied by many since then because it can model numerous phenomena in materials science, such as phase transformations in solids or the thermal expansion problem, as well as being useful as a simple modelling tool for inhomogeneous media [6]

In the case of non-uniform far field conditions, Eshelby showed that if the loading is a polynomial of order n, the interior field is characterized by a polynomial of the same order. This is often called *Eshelby's polynomial conservation theorem*. Mura ([4]) gives further details of this theory and develops a method of solution based on multipole expansions in order to evaluate the induced strain field in an ellipsoidal inhomogeneity using the eigenstrain concept.

In many applications however, heterogeneities are not isolated, since boundary effects or other inhomogeneities influence interior fields. Furthermore inhomogeneities are frequently non-ellipsoidal, particularly for more modern advanced materials applications. Therefore, new improved, predictive methods are required in order to characterize interior fields in composites incorporating inhomogeneities of arbitrary shape for general imposed fields.

Here, a new method is developed in order to approximate the interior fields inside isolated inhomogeneities of arbitrary shape with prescribed polynomial far field conditions.

 ISBN: 978-84-608-6082-2

It is introduced in the context of the two dimensional potential problem (with applications in electrical conductivity and permittivity, magnetic permeability, and elsewhere) but is extendable to three dimensions, and to incorporate interaction and boundary effects. The latter will be discussed in future work. The method begins with the integral equation form of the problem and relies upon series approximations of solutions and the ability to evaluate specific integrals that arise in this context.

To fix ideas the discussion shall focus on the steady state heat conduction problem and the equation governing temperature ϕ in an unbounded homogeneous medium $\Omega \subset \mathbb{R}^2$, in which is embedded a single, isolated inhomogeneity V (for now of general shape and with surface ∂V) satisfying

$$\frac{\partial}{\partial x_j}\left(K_{ij}(\mathbf{x})\frac{\partial \phi}{\partial x_i}\right) = 0 \quad \text{in } \Omega \tag{1}$$

where the assumption is that no heat sources are present. Repeated indices implies summation and since the problem is two dimensional, $\mathbf{x} = (x_1, x_2)$. The inhomogeneous conductivity tensor has components

$$K_{ij}(\mathbf{x}) = K^1_{ij}\chi(\mathbf{x}) + K^0_{ij}(1 - \chi(\mathbf{x}))$$

where χ the characteristic function associated with the inhomogeneity V and K^0_{ij} and K^1_{ij} are the respective uniform conductivity tensors of the host and the inhomogeneity.

The equation can be written in integral form by introducing the associated free-space Green's function of the *host* phase G, defined via the governing equation

$$\frac{\partial}{\partial x_j}\left(K^0_{ij}(\mathbf{x})\frac{\partial G}{\partial x_i}\right) + \delta(\mathbf{x} - \mathbf{y}) = 0, \tag{2}$$

where $\delta(\mathbf{x})$ is the Dirac Delta function, in addition to the far-field condition $\lim_{x\to\infty} G(\mathbf{x}) = 0$. It is then straightforward to show that for every $\mathbf{y} \in V$

$$e_i(\mathbf{y}) = e^\infty_i(\mathbf{y}) + (K^1_{kj} - K^0_{kj})\frac{\partial^2}{\partial y_i \partial y_j}\int_V e_k(\mathbf{x})G(|\mathbf{x} - \mathbf{y}|)\, d\mathbf{x} \tag{3}$$

where the ith component of the temperature gradient has been defined as $e_i = \frac{\partial \phi}{\partial x_i}$ and $e^\infty_i = \frac{\partial \phi^\infty}{\partial x_i}$, with $\phi^\infty(\mathbf{y})$ the solution to the equivalent problem satisfying (1) with no inhomogeneity present. For isotropic materials (3) becomes

$$e_i(\mathbf{y}) = e^\infty_i(\mathbf{y}) + (\kappa_1 - \kappa_0)\frac{1}{2\pi\kappa^0}\frac{\partial^2}{\partial y_i \partial y_j}\int_V e_j(\mathbf{x}) \ln|\mathbf{x} - \mathbf{y}|\, d\mathbf{x} \tag{4}$$

where κ_1 and κ_0 are the thermal conductivities of the inhomogeneity and host regions respectively.

 ISBN: 978-84-608-6082-2

In this work a technique is introduced that can be employed for the prediction of interior fields $e_i(\mathbf{y})$ for general shaped inhomogeneities V. This method is based on the integral equation method introduced in [5] and extended in [3]. More specifically, assuming that the imposed temperature field has the form of a polynomial of order $\mathcal{M}+\mathcal{N}$, i.e.

$$\phi^{\infty}(\mathbf{x}) = \beta_{10}x_1 + \beta_{01}x_2 + \sum_{m=1}^{\mathcal{M}}\sum_{n=1}^{\mathcal{N}} \frac{1}{m!n!}\beta_{mn}x_1^m x_2^n, \quad \beta_{mn} \in \mathbb{R}, \tag{5}$$

the form of the interior field is to be determined. For this purpose the following polynomial expressions within V are posed, of order $M+N$,

$$\phi(\mathbf{x}) = \alpha_{10}x_1 + \alpha_{01}x_2 + \sum_{m=1}^{M}\sum_{n=1}^{N} \frac{1}{m!n!}\alpha_{mn}x_1^m x_2^n, \tag{6}$$

noting that generally $M+N \neq \mathcal{M}+\mathcal{N}$.

A number of restrictions on the form of the constants α_{mn}, β_{mn} already exist since ϕ and ϕ_{∞} are harmonic functions. A linear system for the determination of α_{mn} given β_{mn} arises by substituting (6) and (5) into (4). That this is possible is due to being able to determine forms for the following integral expressions that arise in the governing equation, being denoted by

$$\kappa_0 J^{\delta\xi}(\mathbf{y}) = \frac{1}{2\pi}\int_V x_1^{\delta}x_2^{\xi}\ln|\mathbf{x}-\mathbf{y}|d\mathbf{x}. \tag{7}$$

The so called *generalized Hill tensor* is defined by

$$P_{ij}^{\delta\xi}(\mathbf{y}) = \frac{\partial^2 J^{\delta\xi}(\mathbf{y})}{\partial y_i \partial y_j}, \tag{8}$$

generalized in the sense that the $\delta = \xi = 0$ case refers to the classical Hill tensor [6]. A polynominal approximation to (7) is posed in the form

$$\begin{aligned}\kappa^0 J^{\delta\xi}(\mathbf{y}) &= D_{00}^{\delta\xi} + D_{10}^{\delta\xi}y_1 + D_{01}^{\delta\xi}y_2 + \sum_{p+q=2}^{\mathcal{P}} \frac{D_{pq}^{\delta\xi}}{p!q!}y_1^p y_2^q \\ &= a_0^{\delta\xi}(\rho) + \sum_{m=1}^{\mathcal{P}}\left(a_m^{\delta\xi}(\rho)\cos(m\theta) + b_m^{\delta\xi}(\rho)\sin(m\theta)\right),\end{aligned} \tag{9}$$

where the latter form is determined by writing $\mathbf{x} = \rho(\cos\theta, \sin\theta)$, i.e. in cylindrical polar coordinate form, and $a_m^{\delta\xi}$, $b_m^{\delta\xi}$ are specified in terms of $D_{ij}^{\delta\xi}$ and ρ.

Now exploit orthogonality, setting $\rho = 1$ without loss of generality and applying the following integral operator to (7) and (9) and equating these,

$$\begin{array}{rccl}\mathcal{L}_{(A,m)}: & \mathcal{L}^1([0,2\pi]) & \longmapsto & (-\infty,+\infty) \\ & f & \longmapsto & \displaystyle\int_0^{2\pi} f\cdot(A\cos\, m\theta + (1-A)\sin\, m\theta)\, d\theta.\end{array}$$

 ISBN: 978-84-608-6082-2

with $A = 0, 1$ as required. This permits the determination of $D_{ij}^{\delta\xi}$ for a given V.

Finally therefore, employ the polynomial form (7) together with (6) and (5) in (4) which, upon equating coefficients of each order in the polynomials specifies a linear system for the unknowns α_{ij}. Therefore the temperature ϕ in V (not necessarily elliptic) can be determined for a prescribed ϕ^∞.

In this work the construction above will be implemented, determining the temperature ϕ for a given ϕ^∞ induced on inhomogeneities with several non-elliptical shapes. In particular, this also allows us to confirm Eshelby's polynominal conservation theorem.

Acknowledgements

Parnell is grateful to the Engineering and Physical Sciences Research Council for funding his fellowship (EP/L018039/1). Calvo has been partially supported by the project MTM 2011-24457 of the "Ministerio de Ciencia e Innovación" of Spain and the research group FQM-309 of the "Junta de Andalucía".

References

[1] J.D. Eshelby, The determination of the elastic field of an ellipsoidal inclusion and related problems, *Proc. R. Soc. London A* 241 (1957) 376396.

[2] J. D. Eshelby, Elastic Inclusions and Inhomogeneities, Progress in Solid Mechanics, 2, I. N. Sneddon and R. Hill, eds., North-Holland, Amsterdam, (1961)

[3] D. Joyce, W.J. Parnell, R. Assier and I.D. Abrahams. An integral equation method for the homogenization of unidirectional fibre reinforced media. In preparation.

[4] T. Mura, Micromechanics of Defects in Solids, Martinus Nijhoff Publishers, Dordrecht, 1982.

[5] W.J. Parnell and I.D. Abrahams. A new integral equation approach to elastodynamic homogenization" *Proc. Roy. Soc. A* 464 (2008), 1461-1482

[6] W. J. Parnell, The Eshelby, Hill, Moment and Concentration Tensors for Ellipsoidal Inhomogeneities in the Newtonian Potential Problem and Linear Elastostatics, *J. Elasticity* (2016), 1-64.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Global Optimization-Density Functional Theory Study of Tin Oxide Clusters: Structures, Energies, and Trends

Wesley Paul[1] **and René Fournier**[1]

[1] *Department of Chemistry, York University*

emails: wesleypaul1991@gmail.com, renef@yorku.ca

Abstract

We describe a Global Optimization (GO) strategy in which we search for the minimum energy geometry of several cluster species simultaneously. A relative index of thermodynamic stability is used to compare energies of different species. We show the results for 38 Sn_mO_n cluster species ($m + n \leq 18$). Except for Sn_8O_{10}, the nine most stable species have $m = n$ or $m = n \pm 1$. The global minima structures are open structures with alternating Sn and O atoms. Coordination numbers range between 2 and 7 for Sn atoms and are 3 for most oxygen atoms.

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

1 Introduction

Molecular geometry is a central concept in chemistry. It is hard to make sense of molecular properties or chemical reactivity without knowing where atoms are in a molecule! *Cluster* geometries present particular challenges. In experiments, clusters often exist in a broad range of size and composition and are found in extremely low concentrations. Only a few experimental techniques give structure sensitive information about clusters. Notable among them are photoelectron spectroscopy (PES)[1] and infrared photodissociation spectroscopy (IR-PD)[2, 3]. In recent years, those two techniques have been combined with Density Functional Theory (DFT) to give rich structural information on many cluster species. For instance, there have been PES studies of B_n^-[4], CoB_{16}^-[5], Au_{26}^-[6], $Au_2(AlO)_2$[7], $Ta_2B_6^-$[8],

 ISBN: 978-84-608-6082-2

and metal-centered $M@B_n^-$ wheels[9]; and IR-PD studies of Co_n^+ (n =3–8)[10], Si_nC^+ (n =3–8)[11], $AuSi_n^+$ (n =6–11)[12], Ru_n^+ (n =7–9)[13], and Rh_8^+[14]. DFT results depend, as always, on the exchange-correlation model that is used, on basis sets and/or pseudopotentials, and a number of other approximations (e.g., neglect of spin-orbit coupling) and numerical parameters. *In addition, they depend critically on the quality of the global optimization (GO).*

In the next section we describe a general GO strategy which we combine with DFT to search for the global minima (GM) of clusters. We typically use standard methods of DFT, such as the PBE functional[15] and SDD pseudopotentials[16] with their corresponding Gaussian basis sets[17] to study several cluster species. Our emphasis is on *trends* in structures and energies of neutral clusters as a function of size and composition. We also try to discover, by computation, clusters with unusual stability — so-called "magic clusters".

In the last section we discuss recent results on 38 tin oxide cluster species.

2 Parallel Global Optimization

Computational studies normally look at *groups* of clusters, e.g., M_{13}[18], Sn_n (n = 6–20)[20], Ag_n (n = 4–20)[21], $(MgO)_n$ (n = 3–16)[22]. Typically, GO is performed independently for each species. But clusters within a group are often similar: the GM of Ag_n ($n >$ 6) are all polytetrahedra, none of the $(MgO)_n$ GM has any Mg-Mg or O-O bond, C_n ($n \geq$60) are all cages, etc. So it may be advantageous to optimize all species in a group of clusters concurrently. This is what we do, using an evolutionary strategy similar to a genetic algorithm[23] and operations like "add an atom", "delete an atom", and "transmute atom of element X to element Y", to make child clusters "C" with a different chemical composition than their parents. A good solution for one species can then be used to find good solutions for other species. We also have mating and mutation operations that change the geometry but not the chemical composition, of course. Details can be found elsewhere[24].

Another key feature is the way we implement a memetic algorithm[25]. Every child cluster C obtained by a mating or mutation operation undergoes local optimization *using an auxiliary model potential $F(X)$, not DFT*. This avoids wasting time on geometries with unphysical bond lengths, and avoids the costly DFT minimizations of basin-hopping[19]. In our method, DFT calculations are done only for configurations that are *near* local minima — how near those are to DFT minima depends on the quality of the model potential $F(X)$ — and at the very end, for local optimization of a select few geometries.

Taboo-like penalty terms is another key element of our approach. When cluster "C" is created, we compare it to all previous configurations "X_j" of that species visited so far. If C is too similar to some X_j, we reject it. This avoids repeating costly DFT calculations for the same, or similar, solutions. At the start of a GO run, a modest degree of similarity is enough to reject C. This emphasizes exploration. A progressively tighter similarity criterion is used

 ISBN: 978-84-608-6082-2

to gradually shift emphasis from exploration to exploitation. Near the end, the search is confined near the best solutions, often creating children that differ from their parents by a single bond being made or broken.

3 Sn_mO_n clusters

In bulk, tin oxide exists as stannous oxide (SnO) with 4-coordinate square pyramidal Sn atoms, and stannic oxide (SnO_2) with coordination numbers 6 and 3 for Sn and O, respectively. Small clusters Sn_mO_n could in principle exist in a variety of stochiometric ratios m/n. We wish to find what are the favored species in small Sn_mO_n clusters and uncover structural and energetic trends.

We did two runs of parallel GO, one with $2 \leq m \leq 6$ and $2 \leq n \leq 6$ (25 species) and 6000 iterations in total, the other with $m = 7, 8$ and $4 \leq n \leq 10$ (14 species) and 10000 iterations in total. This was followed by local optimizations for the L best isomers of each species, with L =8 to 12 in all but the smallest species. All calculations were performed with Gaussian09[17], the PBE functional[15], and the SDD pseudopotential[16] and associated basis sets[17]. Sn_2O_4 failed to reach convergence and will be ignored in the following.

The 38 GM, and the 27 isomers found within 1 eV of GM, are open structures with few Sn-Sn or O-O bonds. Figure 1 shows the GM structures of the 9 most stable cluster species. Most oxygen atoms (in grey) have a coordination of 3, but 2 and 4 are also fairly common. The coordination numbers of Sn atoms range from 2 (e.g., Sn_7O_7) to 7 (see Sn_8O_9). Distorted cubes of alternating Sn and O show up in Sn_4O_4 and other clusters (e.g., Sn_7O_7). The GM of Sn_4O_3 (not shown) is a distorted cube with a missing O atom. Distorted Sn_2O_2 squares are a common motif. A few clusters (e.g., Sn_7O_6) have a distorted cage-like structure.

We quantify the degree of short-range mixing with a variable M_6,

$$s = (1/N_s) \sum_{i>j}^{s} (d_{ij}/(R_i + R_j))^{-6} \quad (1)$$

$$d = (1/N_d) \sum_{i>j}^{d} (d_{ij}/(R_i + R_j))^{-6} \quad (2)$$

$$N_s = m(m-1)/2 + n(n-1)/2 \quad (3)$$

$$N_d = mn \quad (4)$$

$$M_6 = d/(d+s) \quad (5)$$

R_i is the atomic radius of the element, N_d is the number of MgO pairs, and the superscript d ("different") on the summation sign indicates a sum over MgO pairs only. Likewise N_s and the superscript s ("same") are for MgMg and OO pairs. We define M_{12} like M_6 but

 ISBN: 978-84-608-6082-2

with -12 as the power. The M_6 of the GM are high, they range from 0.62 to 0.85. On average, the M_6 or low-lying isomers are comparable to those of the GM.

We calculated energies $E(m,n) < 0$, including the zero-point energy, relative to dissociated atoms:

$$m\mathrm{Sn} + n\mathrm{O} \rightarrow \mathrm{Sn}_m\mathrm{O}_n \qquad\qquad E(m,n) \tag{6}$$

In order to make comparisons, we fitted a simple function to the $E(m,n)$

$$E(m,n) \;\approx\; F(m,n) = mE_{Sn} + nE_O \tag{7}$$

where E_{Sn} and E_O, the parameters of the fit, are the negative of mean cohesive energies. We get $E_{Sn} = -3.266\,\mathrm{eV}$ and $E_O = -4.100\,\mathrm{eV}$. Next we define the actual-minus-fit energy difference D, and a dimensionless relative index of thermodynamic stability R[24].

$$D(m,n) \;=\; E(m,n) - F(m,n) \tag{8}$$

$$R(m,n) \;=\; \frac{D(m,n)}{|F(m,n)|} \tag{9}$$

$R(m,n)$ is close to zero on average, negative for cluster species that are stable relative to others, and positive for unstable ones. Table 1 shows $D(m,n)$, $R(m,n)$, M_6, M_{12} and the atomization energy (AE) for the 9 cluster species with smallest R. Their average M_{12} is 0.921 compared to 0.858 for the set of 38 GM. Clearly, highly mixed clusters are energetically favored. It is also clear that cluster species with $m = n$ are energetically favored ($\mathrm{Sn_6O_6}$ ranks 15th in stability out of 38).

We divided the 38 species into four quartiles by rank of their $R(m,n)$, 1 to 9, 10 to 19, 20 to 29, and 30 to 38. Table 2 shows mean values for each quartile.

$\mathrm{Sn}_m\mathrm{O}_n$ clusters get more stable as they get bigger, closer to a 1:1 stochiometric ratio, and better mixed. Those effects are large. For instance, the difference in $D(m,n)$ between $\mathrm{Sn_8O_5}$ and $\mathrm{Sn_7O_6}$ is $1.70 - (-1.79) = 3.49\,\mathrm{eV}$ ($0.27\,\mathrm{eV/atom}$). To further quantify this, we define $N = m + n$, the composition difference $c_d = |m - n|$, and the composition ratio $c_r = Max(m/n, n/m)$. The result of linear regressions of $R(m,n)$ with respect to those variables and M_6 are summarized in Table 3. The clearest trends are increasing stability (smaller R) as c_r decreases and approaches 1 ($\mathrm{Sn}_m\mathrm{O}_m$ species), and increasing mixing index as c_r decreases (see Figure 2). We rationalize this as a finite-size effect. In small $\mathrm{Sn}_m\mathrm{O}_n$ clusters the only geometries with large M_6 that satisfy basic constraints have $m \approx n$. In particular, $(\mathrm{SnO_2})_n$ can not be well mixed when n is small. Since ionic bonding increases with mixing, $m \approx n$ clusters are more stable. In the bulk limit, there *is* a well-mixed geometry with $n/m = 2$ — the rutile structure of $\mathrm{SnO_2}$. It would be interesting to find out the smallest n at which $(\mathrm{SnO_2})_n$ becomes stable.

`http://pirweb.edv.uniovi.es/webcmmse/index.php`

 ISBN: 978-84-608-6082-2

Table 1: Stability indices D and R, mixing indices M_6 and M_{12}, and atomization energy (AE, in eV) for the nine most stable cluster species.

	m.n	$D(m,n)$ (eV)	$R(m,n)$	M_6	M_{12}	AE (eV)
1	7.7	−3.58	−0.0695	0.827	0.952	55.14
2	8.9	−3.94	−0.0626	0.826	0.939	66.97
3	8.7	−2.89	−0.0525	0.772	0.902	57.70
4	7.8	−2.89	−0.0519	0.830	0.939	58.55
5	8.10	−3.43	−0.0511	0.818	0.912	70.55
6	5.5	−1.75	−0.0477	0.821	0.943	38.58
7	4.4	−1.37	−0.0464	0.784	0.907	30.83
8	8.8	−2.34	−0.0397	0.791	0.918	61.26
9	7.6	−1.79	−0.0378	0.759	0.881	49.25

Table 2: Stability index (R), mixing index (M_6), size ($m+n$), and $|m-n|$ for the 38 global minima divided in quartiles using R.

quartile	Mean(R)	Mean(M_6)	Mean($m+n$)	Mean($\lvert m-n\rvert$)
1	−0.051	0.803	14.0	0.67
2	−0.010	0.783	12.1	1.50
3	+0.056	0.740	8.7	1.90
4	+0.152	0.677	7.0	2.56

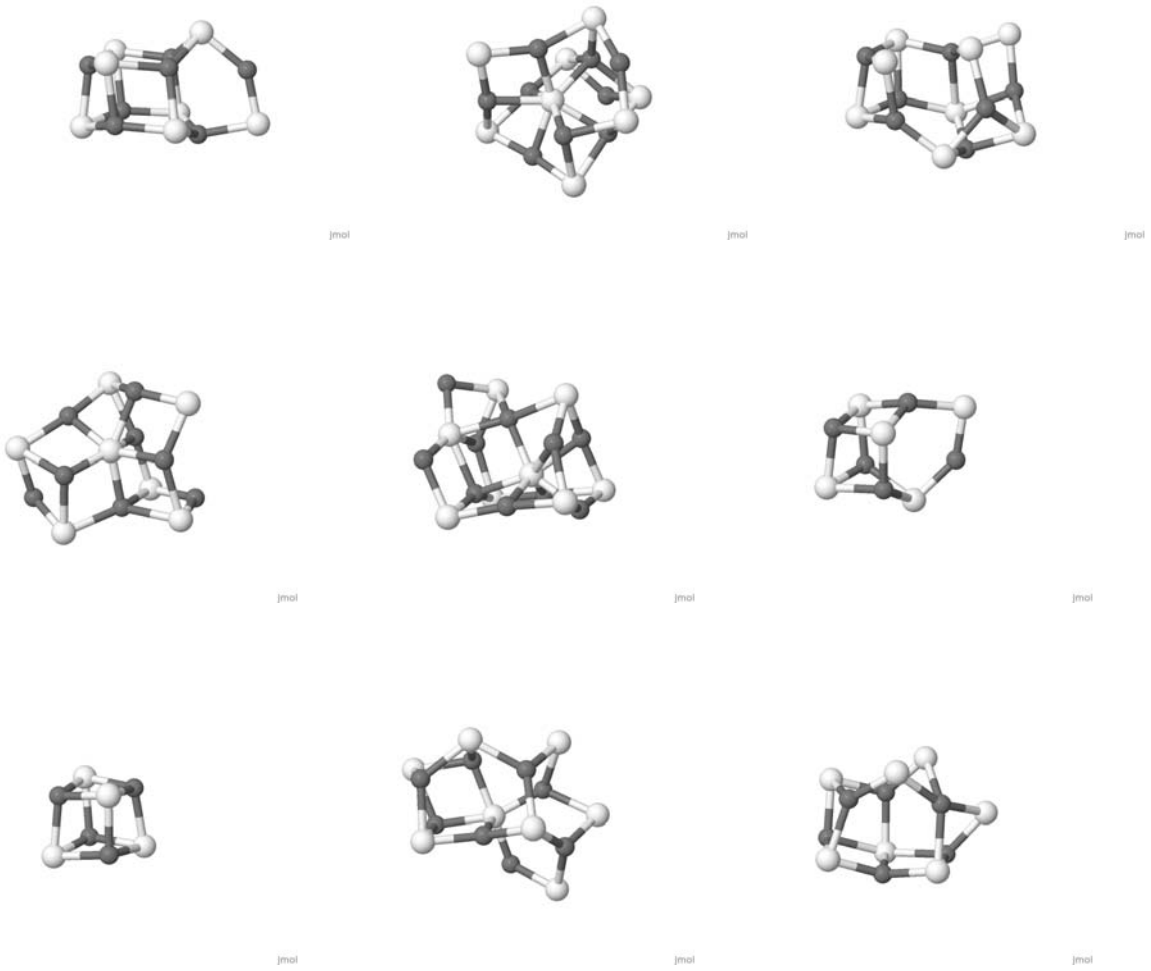
Table 3: Linear regressions of the stability index R vs N, c_d, M_6 and c_r, and of M_6 vs c_r.

	intercept	slope	correlation coefficient
R vs N	0.192	−0.015	−0.70
R vs c_d	−0.026	0.038	+0.57
R vs M_6	0.624	−0.781	+0.67
R vs c_r	−0.151	0.123	+0.82
M_6 vs c_r	0.895	-0.095	−0.74

 ISBN: 978-84-608-6082-2

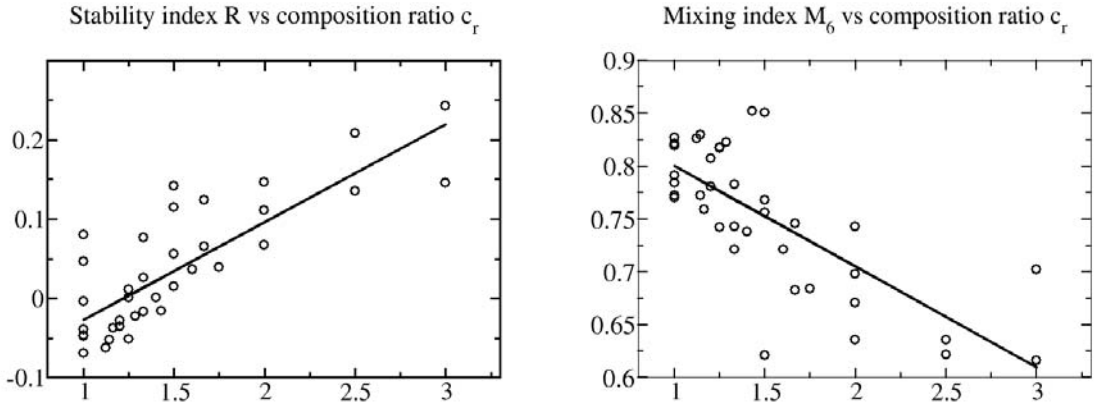
4 Figures

Fig. 1. Global minima structures of the nine most stable Sn_mO_n cluster species, $m.n$ = 7.7, 8.9, 8.7, 7.8, 8.10, 5.5, 4.4, 8.8, and 7.6 (Sn=white, O=grey).



 ISBN: 978-84-608-6082-2

Fig. 2. Correlations between stability index R, mixing index M_6 and composition ratio c_r.



Acknowledgements

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

References

[1] I. Leon, Z. Yang, H. T. Liu and L. S. Wang, *The Design and Construction of A High-Resolution Velocity-Map Imaging Apparatus for Photoelectron Spectroscopy Studies of Size-Selected Clusters*, Rev. Sci. Instrum. **85** (2014) 083196.

[2] M. A. Duncan, *Infrared Spectroscopy to Probe Structure and Dynamics in Metal Ion-Molecule Complexes*, Int. Rev. Phys. Chem. **22** (2003) 407–435.

[3] A. Fielicke, A. Kirilyuk, C. Ratsch, J. Behler, M. Scheffler, G. von Helden, and G. Meijer, *Structure Determination of Isolated Metal Clusters via Far-Infrared Spectroscopy*, Phys. Rev. Lett. **93** (2004) 023401.

[4] L. S. Wang, *Photoelectron Spectroscopy of Size-Selected Boron Clusters: From Planar Structures to Borophenes and Borospherenes*, Int. Rev. Phys. Chem. **35** (2016) 69–142.

[5] I. A. Popov, T. Jian, G. V. Lopez, A. I. Boldyrev and L. S. Wang, *Cobalt-Centered Boron Molecular Drums with the Highest Coordination Number in the CoB_{16} Cluster*, Nat. Commun. **6** (2015) 8654.

[6] B. Schaefer, R. Pal, N. S. Khetrapal, M. Amsler, A. Sadeghi, V. Blum, X. C. Zeng, S. Goedecker and L. S. Wang, *Isomerism and structural fluxionality in the Au_{26} and Au_{26}^- nanoclusters*, ACS Nano **7** (2014) 7413–7422.

[7] G. V. Lopez, T. Tian, W.-L. Li and L. S. Wang, *Electronic structure and chemical bonding of a highly stable and aromatic auro-aluminum oxide cluster*, J. Phys. Chem. A **118** (2014) 5204–5211.

[8] W.-L. Li, L. Xie, T. Jian, C. Romanescu, X. Huang and L. S. Wang, *Hexagonal bipyramidal $Ta_2B_6^{-/0}$ clusters: B_6 rings as structural motifs*, Angew. Chem. **126** (2014) 1312–1316

[9] C. Romanescu, T. R. Galeev, W.-L. Li, A. I. Boldyrev and L.S. Wang, *Transition-metal centered monocyclic boron wheel clusters $M@B_n$: a new class of aromatic borometallic compounds*, Acc. Chem. Res. **46** (2012) 350–358.

[10] R. Gehrke, P. Gruene, A. Fielicke, G. Meijer, K. Reuter, *Nature of Ar Bonding to Small Co_n^+ Clusters and its Effect on the Structure Determination by Far-Infrared Absorption Spectroscopy*, J. Chem. Phys. **130** (2009) 034306.

[11] N. X. Truong, M. Savoca, D. J. Harding, A. Fielicke, and O. Dopfer, *Vibrational spectra and structures of Si_nC clusters (n =3–8)*, Phys. Chem. Chem. Phys. **17** (2015) 18961-18970.

[12] Y. Li, J. T. Lyon, A. P. Woodham, P. Lievens, A. Fielicke, and E. Janssens, *Structural Identification of Gold-Doped Silicon Clusters via Far-Infrared Spectroscopy*, J. Phys. Chem. C **119** (2015) 10896–10903.

[13] C. Kerpal, D. J. Harding, D. M. Rayner, J. T. Lyon, and A. Fielicke, *Far-IR Spectra and Structures of Small Cationic Ruthenium Clusters: Evidence for Cubic Motifs*, J. Phys. Chem. C **119** (2015) 10869–10875.

[14] D. J. Harding, T. R. Walsh, S. M. Hamilton, W. S. Hopkins, S. R. Mckenzie, P. Gruene, M. Haertelt, G. Meijer, and A. Fielicke, *The structure of Rh_8^+ in the gas phase*, J. Chem. Phys. **132** (2010) 011101.

[15] J. P. Perdew, K. Burke, and M. Ernzerhof, *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77** (1996) 3865–3868.

[16] A. Bergner, M. Dolg, W. Kuechle, H. Stoll, and H. Preuss, *Ab-initio energy-adjusted pseudopotentials for elements of groups 13-17*, Mol. Phys. **80** (1993) 1431–1441.

 ISBN: 978-84-608-6082-2

[17] Gaussian 09, Revision E.01, M. J. Frisch et al., Gaussian, Inc., Wallingford CT, 2009.

[18] Y. Sun, M. Zhang and R. Fournier, *Periodic Trends in the Geometric Structures of 13-Atom Metal Clusters*, Phys. Rev. B **77** (2008) 075435.

[19] D. J. Wales and J. P. K. Doye, *Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 Atoms*, J. Phys. Chem A **101** (1997) 5111–5116.

[20] S. Schäfer, B. Assadollahzadeh, M. Mehring, P. Schwerdtfeger, and R. Schäfer, *Structure and electric properties of Sn_n clusters (n =6–20) from combined electric deflection experiments and quantum theoretical studies*, J. Phys. Chem. A **112** (2008) 12312–12319.

[21] H. Dhillon and R. Fournier, *Geometric structure of silver clusters with and without adsorbed Cl and Hg*, Comp. Theor. Chem. **1021** (2013) 26–34 (2013).

[22] M. Haertelt, A. Fielicke, G. Meijer, K. Kwapien, M. Sierka and J. Sauer, *Structural determination of neutral MgO clusters — hexagonal nanotubes and cages*, Phys. Chem. Chem. Phys. **14** (2012) 2849–2856.

[23] J. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.

[24] R. Fournier and A. Mohareb, *Optimizing Molecular Properties Using a Relative Index of Thermodynamic Stability and Global Optimization Techniques*, J. Chem. Phys. **144** (2016) 024114.

[25] P. Moscato, *On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms: Caltech Concurrent Computation Program (report 826)* (1989)

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Stabilization of Switched Linear Systems by using Projections

C. Pérez[1], F. Benítez-Trujillo[1] and J. B. García-Gutiérrez[1]

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

emails: carmen.perez@uca.es, quico.benitez@uca.es,
juanbosco.garciagutierrez@alum.uca.es

Abstract

This paper proposes a new approach to stabilize switched linear systems. In this method, the projections are employed to establish the stabilization of a switched systems class. Hence, we suppose that the subsystems given by a switched linear system can be projected to the same subspace. Under these conditions, we prove that this switched linear system is stabilizable if and only if a low-order switched linear system is also stabilizable. In order to complete this study, we present a counter example that proves that it is not always possible to use the projections. Moreover, the main result in the paper is applied to solve the stabilization of a third-order switched systems class. Finally, a numerical example is included in order to illustrate the new method obtained to stabilize a third-order switched systems class.

Key words: Projections, Stabilization, Switched systems, Conic switching law

1 Introduction

In recent years, the study of switched systems has received more and more attention; see, e.g., [6], [7], [11], among many others. This is because switched systems have strong applications in Control Theory. One of the main issues in the study of switched systems is stability.

The problems of stability of switched systems include two aspects: one is how to make switched systems stable or stabilized under arbitrary switching law, where each switching subsystem is required to be stable (see [6], [8], [12]). One of the main results is obtained in [5], where it is shown that the switched linear system is exponentially stable under arbitrary switching laws if the Lie algebra generated by the subsystems is solvable. In particular, this

implies that the matrices can be written in a triangular form, then the system is stable under arbitrary switching. However, for the nonlinear case, as demonstrated in [1], the triangular structure alone is not sufficient for stability.

The second aspect of the stability is how to design a switching law under which switched systems are stable or stabilized, where each switched subsystem is unstable or unstabilizable. We can find some conditions that guarantee the existence of a *stabilizing* switching law ([4], [13], [14]). For second-order switched systems, methods for stabilizing switched systems have been presented ([2], [3], [15]). In [2] and [15] there are necessary and sufficient conditions that solve the problem of stabilization of switched systems consisting of unstable second-order linear subsystems. Furthermore, the papers [3] and [9] employ these ideas to study the nonlinear case and solve the linear case for two subsystems with different equilibrium types.

In this paper, we present a new result about the stabilization of a switched systems class. This new result establishes that the stabilization of these switched systems is equivalent to the stabilization of a low-order switched system. Hence, in this case, the problem of stabilization is reduced to study switched systems in a lower dimension.

As an application of this result, the results in [2] and [15] are used to study the stabilization of higher order switched systems by projecting the trajectory of the system to some 2-dimensional subspaces. We find these subspaces and establish conditions for their existence.

The remainder of this paper is organized as follows. In Section 2 the problem is formulated and a preliminary result is enunciated. Section 3 deals with the main result aforementioned. In Section 4 we present a counter example of the extension of the main result to another switched linear systems class. In Section 5 the application of the result in Section 3 to the stabilization of a third-order switched systems class is provided. Finally, in Section 6 we provide the conclusions.

2 Preliminaries

In this paper, given a family of real $n \times n$ matrices $\{A_p : p \in \mathcal{P}\}$, where the index set $\mathcal{P}$ is an arbitrary compact set, switched linear systems of the following form are studied

$$\dot{x}(t) = A_{\sigma(t)}x(t), \tag{1}$$

where $x \in \mathbb{R}^n$ and $\sigma : [0, \infty) \longrightarrow \mathcal{P}$ is a piecewise constant function called *switching law* indicating the active subsystem at each instant of time.

Given a linear subspace W of $\mathbb{R}^n$, we will say that W is an invariant subspace of the family of matrices $\{A_p : p \in \mathcal{P}\}$ if $A_pW \subset W$ for all $p \in \mathcal{P}$. The following proposition will be used later.

Proposition 1 *If W is a linear subspace of $\mathbb{R}^n$ of dimension $m \leq n$ and $\{v_1, v_2, \ldots, v_m\}$ is a basis of W, there exist $n-m$ vectors, $e_{m+1}, e_{m+2}, \ldots, e_n$ such that*

$$\{v_1, \ldots, v_m, e_{m+1}, \ldots, e_n\}$$

is a basis in $\mathbb{R}^n$.

Hence, if W is an invariant subspace of the previous family of matrices, applying the previous proposition and defining the matrix

$$P = [v_1, \ldots, v_m, e_{m+1}, \ldots, e_n],$$

it is obtained that

$$P^{-1} A_p P = \begin{bmatrix} B_p & C_p \\ \overline{0} & D_p \end{bmatrix},$$

where B_p is a $m \times m$ matrix, C_p is a $m \times (n-m)$ matrix, $\overline{0}$ is the $(n-m) \times m$ matrix identically zero, and D_p is a $(n-m) \times (n-m)$ matrix. Then, we have obtained two new families of square matrices, $\{B_p : p \in \mathcal{P}\}$ and $\{D_p : p \in \mathcal{P}\}$.

For notational simplicity, we introduce the linear function $p_{n-m} : \mathbb{R}^n \longrightarrow \mathbb{R}^{n-m}$ that assigns to each $x \in \mathbb{R}^n$ the last $n-m$ coordinates of x in the basis given by $\{v_1, \ldots, e_n\}$, i.e., $p_{n-m}(x_1, x_2, \ldots, x_n) = (x_{m+1}, \ldots, x_n)$ where $(x_1, x_2, \ldots, x_n)$ are the coordinates of x in this basis.

Before proceeding to the next section, we introduce some definitions about stability of a system. Consider the system given by

$$\dot{x} = f(x, t), \qquad x(t_0) = x_0, \tag{2}$$

where $x \in \mathbb{R}^n$ and $t \geq 0$. If we assume that f is *piecewise* continuous with respect to t, i.e., there are only finitely many discontinuity points in any compact set and $f(0, t) = 0$ for any $t \geq 0$, we can define the following (see more details in [10]).

Definition 1 *The equilibrium point $x = 0$ is a locally exponentially stable equilibrium point of (2) if there exist $k, \alpha > 0$ such that*

$$\|x(t)\| < k e^{-\alpha(t-t_0)} \|x(t_0)\|$$

for all x_0 in a neighborhood of the origin and $t \geq t_0 \geq 0$.

Global exponential stability is defined by requiring the previous inequality to hold for all $x_0 \in \mathbb{R}^n$. We always consider systems for which the origin is an equilibrium point and, with some abuse, we say that a system is exponentially stable, meaning that the origin is an exponentially stable equilibrium point of the system.

3 Main result

Under the previous notation, we present the main result in this work.

Theorem 1 *Let $\{A_p : p \in \mathcal{P}\}$ be a family of $n \times n$ matrices where $\mathcal{P}$ is compact. Suppose that there exists W an invariant subspace of this family of matrices. Thus, under the previous notation, if B_p is a stable matrix, for all $p \in \mathcal{P}$, and $\dot{y}(t) = B_{\sigma(t)}y$ is exponentially stable under a switching law σ, we have that:*

The switched system $\dot{x} = A_\sigma x$ is exponentially stable under the switching law σ if and only if the switched system $\dot{z} = D_\sigma z$ is also exponentially stable under the same switching law σ.

Proof. We can assume that, by a linear change of coordinates, the matrices A_p are given by

$$A_p = \begin{bmatrix} B_p & C_p \\ \overline{0} & D_p \end{bmatrix}.$$

Firstly, we will suppose that the switched system $\dot{x} = A_\sigma x$ is globally exponentially stable. Consider the switched system $\dot{z}(t) = D_{\sigma(t)}z(t)$ under the switching law σ. In order to prove that this system is globally exponentially stable, we pick $z_0 \in \mathbb{R}^{n-m}$ and consider the solution of $\dot{x}(t) = A_{\sigma(t)}x(t)$ with $x(0) = (0, z_0)$ where 0 is the null vector in $\mathbb{R}^m$.

As this system is globally exponentially stable, there exist k, $\alpha > 0$ such that

$$\|x(t)\| < ke^{-\alpha t}\|x(0)\|.$$

If we denote $x(t) = (y(t), z(t))$ where $z(t) = p_{n-m}(x(t))$, it is obtained that

$$\dot{x}(t) = \begin{bmatrix} \dot{y}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} B_p & C_p \\ \overline{0} & D_p \end{bmatrix} \begin{bmatrix} y(t) \\ z(t) \end{bmatrix} \tag{3}$$

Therefore, $\dot{z}(t) = D_{\sigma(t)}z(t)$ and $z(0) = z_0$, and this implies that $z = p_{n-m} \circ x$ is the solution of the switched system given by the family $\{D_1, D_2, \ldots, D_N\}$ under the switching law σ and if P_{n-m} is the matrix given by the linear function p_{n-m}, we have that

$$\|z(t)\| = \|Mx(t)\| \leq \|P_{n-m}\| \cdot \|x(t)\| < \|P_{n-m}\| \cdot ke^{-\alpha t}\|x_0\|.$$

As $\|x_0\| = \|z_0\|$, it holds that

$$\|z(t)\| < \|P_{n-m}\| \cdot ke^{-\alpha t}\|z_0\|.$$

Therefore the switched system $\dot{z} = D_\sigma z$ is globally exponentially stable.

Now, we will suppose that the switched system $\dot{z} = D_\sigma z$ is globally exponentially stable under some switching law σ. Therefore, there are c, $\mu > 0$ such that the solution of $\dot{z} = D_\sigma z$ for every $z_0 \in \mathbb{R}^{n-m}$ verifies

$$\|z(t)\| < ce^{-\mu t}\|z_0\|. \tag{4}$$

In order to prove that $\dot{x} = A_\sigma x$ is globally exponentially stable, we choose the initial condition $x_0 \in \mathbb{R}^n$. If x is the solution of this switched system and we write x as $x(t) = (y(t), z(t))$ where $z = p_{n-m} \circ x$, then it is satisfied (3) and, thus, $z = p_{n-m} \circ x$ is the solution of $\dot{z} = D_\sigma z$ for the initial condition $z(0) = z_0$ con $z_0 = p_{n-m}(x_0)$ and y is the solution of the following switched linear control system

$$\begin{aligned} \dot{y}(t) &= B_{\sigma(t)} y(t) + C_{\sigma(t)} z(t) \\ y(0) &= y_0 \end{aligned}$$

Given the initial state $y(0) = y_0$ and an interval $[0, T]$ where the switching law σ is defined, if $0 < t_1 < \ldots < t_{N-1} < t_M = T$ is the switching time sequence of σ in $[0, T]$, $h_k = t_k - t_{k-1}$ for $k = 1, 2, \ldots, N$ and

$$\sigma(t_{k-1}) = i_k \in \{1, 2, \ldots, N\}, \text{ for } k = 1, 2, \ldots, N,$$

then the solution of this control system is given by

$$\begin{aligned} y(T) &= \prod_{m=N}^{1} e^{B_{i_m} h_m} y_0 + \sum_{k=1}^{N-1} \left(\prod_{m=N}^{k+1} e^{B_{i_m} h_m} \right) \int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k - s)} C_{i_k} z(s) ds + \\ &+ \int_{t_{N-1}}^{T} e^{B_{i_N}(T-s)} C_{i_N} z(s) ds \end{aligned}$$

As $\dot{\overline{y}}(t) = B_{\sigma(t)} \overline{y}(t)$ is exponentially stable under the switching law σ, there are $\overline{l}$, $\overline{\lambda} > 0$ such that the solution of $\dot{\overline{y}} = B_\sigma \overline{y}$ verifies:

$$\|\overline{y}(t)\| < \overline{l} e^{-\overline{\lambda}(t - t_0)} \|\overline{y}(t_0)\|.$$

Note that, under this notation, with $y(0) = y_0$, this solution is given by

$$\overline{y}(T) = \prod_{m=N}^{1} e^{B_{i_m} h_m} y_0.$$

Thus, we have that $\left\| \prod_{m=N}^{1} e^{B_{i_m} h_m} y_0 \right\| < \overline{l} e^{-\overline{\lambda} T} \|y_0\|$.

Moreover, if we choose as initial condition $\overline{y}(t_k) = \int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k - s)} C_{i_k} z(s) ds$, we also have that:

$$\left\| \left(\prod_{m=N}^{k+1} e^{B_{i_m} h_m} \right) \int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k - s)} C_{i_k} z(s) ds \right\| \leq \overline{l} e^{\overline{\lambda}(T - t_k)} \left\| \int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k - s)} C_{i_k} z(s) ds \right\|$$

 ISBN: 978-84-608-6082-2

Then,

$$\begin{aligned}
&\left\|\sum_{k=1}^{N-1}\left(\prod_{m=N}^{k+1} e^{B_{i_m}h_m}\right)\int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k-s)}C_{i_k}z(s)ds\right\| \\
&\leq \sum_{k=1}^{N-1} \overline{l}e^{-\overline{\lambda}(T-t_k)}\left\|\int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k-s)}C_{i_k}z(s)ds\right\| \\
&\leq \sum_{k=1}^{N-1} \overline{l}e^{-\overline{\lambda}(T-t_k)}\int_{t_{k-1}}^{t_k} \|e^{B_{i_k}(t_k-s)}\|\cdot\|C_{i_k}\|\cdot\|z(s)\|ds
\end{aligned}$$

By (4), we have that $\|z(s)\| < c\cdot e^{-\mu s}\|z_0\|$ and, by hypothesis, for each $p\in\mathcal{P}$, B_p is a stable matrix, then, there are l_p, $\lambda_p > 0$ such that

$$\|e^{B_p(t-t_0)}\| < l_p e^{-\lambda_p(t-t_0)} \qquad \text{for any } t_0 \geq 0.$$

Therefore, by defining $l = \max\{l_p : p\in\mathcal{P}\}$, $\lambda = \min\{\mu, \overline{\lambda}, \lambda_p : p\in\mathcal{P}\}$ and $\overline{C} = \max\{\|C_p\| : p\in\mathcal{P}\}$, it holds that:

$$\|e^{B_{i_k}(t_k-s)}\| \leq le^{-\lambda(t_k-s)} \qquad \text{for each } k = 1,2,\ldots,N-1.$$

Then, by applying this to the previous inequality, it holds that

$$\begin{aligned}
&\left\|\sum_{k=1}^{N-1}\left(\prod_{m=N}^{k+1} e^{B_{i_m}h_m}\right)\int_{t_{k-1}}^{t_k} e^{B_{i_k}(t_k-s)}C_{i_k}z(s)ds\right\| \\
&\leq \sum_{k=1}^{N-1} \overline{l}e^{-\overline{\lambda}(T-t_k)}\int_{t_{k-1}}^{t_k} l\cdot e^{-\lambda(t_k-s)}\cdot\overline{C}\cdot c\cdot e^{-\mu s}\cdot\|z_0\|ds \\
&\leq \sum_{k=1}^{N-1} \overline{l}\cdot l\cdot\overline{C}\cdot c\cdot e^{-\overline{\lambda}(T-t_k)}\int_{t_{k-1}}^{t_k} e^{-\lambda(t_k-s)}\cdot e^{\lambda s}\cdot\|z_0\|ds \\
&= \sum_{k=1}^{N-1} \overline{l}\cdot l\cdot\overline{C}\cdot c\cdot e^{-\overline{\lambda}(T-t_k)}\int_{t_{k-1}}^{t_k} e^{-\lambda t_k}\cdot\|z_0\|ds \\
&= \sum_{k=1}^{N-1} \overline{l}\cdot l\cdot\overline{C}\cdot ce^{-\overline{\lambda}(T-t_k)}h_k e^{-\lambda t_k}\|z_0\| = \overline{l}\cdot l\cdot\overline{C}\cdot c\cdot t_{N-1}\cdot e^{-\lambda T}\|z_0\|,
\end{aligned}$$

so $\mu\geq\lambda$ and $\overline{\lambda}\geq\lambda$.

By applying again that $\|e^{B_{i_N}(T-s)}\| \leq le^{-\lambda_{i_N}(T-s)}$, $\|C_{i_N}\| < \overline{C}$, and $\|z(s)\| < ce^{-\mu s}\|z_0\|$,

$$\left\|\int_{t_{N-1}}^{T} e^{B_{i_N}(T-s)}C_{i_N}z(s)ds\right\| \leq \int_{t_{N-1}}^{T} \|e^{B_{i_N}(T-s)}\|\cdot\|C_{i_N}\|\cdot\|z(s)\|ds$$

$$\leq \int_{t_{N-1}}^{T} l \cdot e^{-\lambda_{i_N}(T-s)} \cdot \overline{C} \cdot c \cdot e^{-\mu s} \cdot \|z_0\| ds \leq l \cdot \overline{C} \cdot c \int_{T_{N-1}}^{T} e^{-\lambda(T-s)} e^{-\lambda s} \cdot \|z_0\| ds$$
$$= l \cdot \overline{C} \cdot c \cdot e^{-\lambda T} \|z_0\| (T - t_{N-1})$$

Therefore, we have obtained that:

$$\|y(T)\| \leq l e^{-\overline{\lambda} T} \|y_0\| + \overline{l} \cdot l \cdot \overline{C} \cdot c \cdot t_{N-1} \cdot e^{-\lambda T} \|z_0\| + l \cdot \overline{C} \cdot c \cdot e^{-\lambda T} (T - t_{N-1}) \|z_0\|.$$

If we define $\tilde{l} = \max\{\overline{l} \cdot l \cdot \overline{C} \cdot c, l \cdot \overline{C} \cdot c\}$, we have that:

$$\|y(T)\| \leq l e^{-\overline{\lambda} T} \|y_0\| + \tilde{l} T e^{-\lambda T} \|z_0\|.$$

Now, it is easy to prove that an exponential function $k' e^{-\gamma T}$ can be found where γ, $k' > 0$ such that

$$\|y(T)\| < k' e^{-\gamma T} \left(\|y_0\| + \|p_{n-m}(x_0)\| \right)$$

Consequently, if we consider the norm $\|\cdot\|_1$, this can be written as

$$\|y(T)\|_1 < k' e^{-\gamma T} (\|y_0\|_1 + \|p_{n-m}(x_0)\|_1) = k' e^{-\gamma T} \|x_0\|_1.$$

Or, likewise,

$$\|y(T)\|_\infty < k e^{\gamma T} \|x_0\|_\infty, \tag{5}$$

so on a finite dimensional vector space all norms are equivalent.

Furthermore, from (4) we know that

$$\|z(T)\|_\infty < c e^{-\mu T} \|z_0\|_\infty \leq c e^{-\mu T} \|x_0\|_\infty. \tag{6}$$

To conclude, if we define $\overline{k} = \max\{c, k'\}$ and $\overline{\gamma} = \min\{\mu, \gamma\}$, then from (5) and (6) it can be deduced that:

$$\|x(T)\|_\infty < \overline{k} e^{-\overline{\gamma} T} \|x_0\|_\infty$$

In summary, by definition, the switched system $\dot{x} = A_\sigma x$ under this switching law σ is exponentially stable. □

4 A counter example

In Theorem 1, we suppose that the matrix B_p, for each $p \in \mathcal{P}$, is stable and the switched system consisting of these matrices is globally exponentially stable under a switching law σ. It seems natural to ask what happens when not all the matrices B_p are stable. Hence, the question is: if there exists a switching law σ such that the switched systems consisting of B_p and D_p are stable, then, the original switched system is also stable under the same switching law?. The answer is no. In order to prove it, we present the following counter example:

 ISBN: 978-84-608-6082-2

Example 1 *Consider the switched system (1) consisting of the following matrices:*

$$A_1 = \begin{bmatrix} \frac{1}{10} & -1 & -2 & 3 \\ 2 & \frac{1}{10} & 4 & 1 \\ 0 & 0 & \frac{1}{10} & -1 \\ 0 & 0 & 2 & \frac{1}{10} \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} \frac{1}{10} & -2 & 1 & 5 \\ 1 & \frac{1}{10} & 4 & 5 \\ 0 & 0 & \frac{1}{10} & -2 \\ 0 & 0 & 1 & \frac{1}{10} \end{bmatrix}$$

Now, if we define W as the linear subspace generated by $e_1 = (1,0,0,0)$ and $e_2 = (0,1,0,0)$, it is clear that W is an invariant subspace of both matrices. Hence, by following the previous notation, we can define the following matrices:

$$B_1 = D_1 = \begin{bmatrix} \frac{1}{10} & -1 \\ 2 & \frac{1}{10} \end{bmatrix} \text{ and } B_2 = D_2 = \begin{bmatrix} \frac{1}{10} & -2 \\ 1 & \frac{1}{10} \end{bmatrix}$$

As $B_i = D_i$, for $i = 1, 2$, we only have to consider the switched system consisting of B_1 and B_2. It is important to note that the matrices B_i, for $i = 1, 2$, are not stable.

Now, if we consider the switching law defined as follows

$$\sigma(t) = \begin{cases} 1 & \text{if } x_1(t)x_2(t) > 0 \\ 2 & \text{if } x_1(t)x_2(t) \leq 0 \end{cases},$$

then, it can be proved that the switched system consisting of B_1 and B_2 under this switching law is globally exponentially stable. In particular, given $x_0 = (1,1,1,1)$, if we consider the initial condition $(1,1)$ for the switched system given by B_1 and B_2, we have that the solution of this switched system converges to the origin (see Figure 1 (a)).

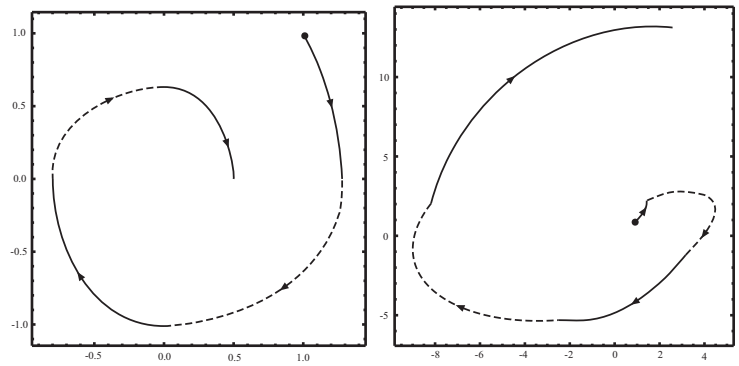
However, for the initial condition x_0 it is easy to check that the solution of the switched system consisting of A_1 and A_2 does not converge to the origin under the switching law σ. In Figure 1 (b) we can see the first and second coordinates of this solution and how they go outwards from the origin.

5 Application to the stabilization of third-order switched systems

In this section, Theorem 1 is applied to establish a sufficient condition for the stabilization of a third-order switched systems class. Consider the switched system (1) given by two non-stable 3×3 matrices, A_1 and A_2, and suppose that there exists an one-dimensional invariant subspace W of A_1 and A_2. As the dimension of W is one, this condition is equivalent to the existence of an eigenvector v common to these matrices. Moreover, the eigenvalues associated to this eigenvector are real since W is one-dimensional.

Using the notation in the "Preliminaries" Section, these matrices can be written in the following form

$$A_i = \begin{bmatrix} B_i & C_i \\ \overline{0} & D_i \end{bmatrix},$$



(a) (b)

Figure 1: (a) Stable trajectory for the switched system given by B_1 and B_2 (or D_1 and D_2) in Example 1. (b) The first and second coordinates of the switched system given by A_1 and A_2 in Example 1.

where $B_i = (\lambda_i)$ with λ_i the eigenvalue associated to the eigenvector v, $\overline{0}^1$ is the vector identically zero for $\mathbb{R}^2$ and D_i is a 2×2 matrix for $i = 1, 2$.

In order to apply Theorem 1, we must suppose that B_i is stable, i.e., λ_i is real and negative, for $i = 1, 2$.

Now, the method for stabilizing second-order switched systems [2] can be applied to the switched system consisting of D_1 and D_2. Hence, if this system is stabilizable, i.e., if it is possible to find a switching law such that the system is stable, then, by Theorem 1, the switched system (1) is stable under the same switching law.

Theorem 2 *Consider the switched system (1) given by two 3×3 non-stable matrices, A_1 and A_2, and suppose that these matrices have a common eigenvector v whose associated eigenvalues, λ_1 and λ_2, are real and negative.*

The switched system (1) is globally exponentially stable under a switching law σ if and only if the switched system given by the matrices D_1 and D_2 is globally exponentially stable under the same switching law.

The proof is omitted because it is deduced from the previous results.

Nowe, in order to illustrate the above development, the following example will be studied.

Example 2 *Consider the switched system (1) consisting of the following matrices*

$$A = \begin{bmatrix} 2 & -13 & 8 \\ 5 & 0 & 2 \\ 5 & 1 & 1 \end{bmatrix} \quad \textit{and} \quad B = \begin{bmatrix} -2 & -48 & 24 \\ 1 & 7 & -4 \\ 1 & 7 & -5 \end{bmatrix}.$$

These matrices have a common eigenvector, $v = (-1, 1, 2)$, associated to the negative real eigenvalues -1 and -2, respectively. Hence, a matrix P can be defined as follows

$$P = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}.$$

Then, we can obtain a equivalent switched system consisting of the following matrices

$$A_1 = P^{-1}AP = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 5 \\ 0 & -5 & 2 \end{bmatrix} \quad \textit{and} \quad A_2 = \begin{bmatrix} -2 & -1 & 0 \\ 0 & 4 & 1 \\ 0 & -25 & -2 \end{bmatrix}$$

Hence, in order to apply the previous theorem, we study the stabilization of the second-order switched systems given by

$$D_1 = \begin{bmatrix} 2 & 5 \\ -5 & 2 \end{bmatrix} \quad \textit{and} \quad D_2 = \begin{bmatrix} 4 & 1 \\ -25 & -2 \end{bmatrix}.$$

By following the results in [15] or [2], it is obtained that the second-order switched system is globally exponentially stable. Consequently, from Theorem 2 it is deduced that the switched system given by A and B is stabilizable (see Figure 2).

6 Conclusions

A new result based on projections for stabilizing switched linear systems has been presented in this paper. With this result, the stabilization of a switched systems class is assured by studying a low-order switched system.

Moreover, we have solved the problem of the stabilization of a third-order switched systems class by using projections and the results of stabilizing second-order switched systems in [15]. Hence, this also provides a partial solution to the open problem posed in [15]; i.e., studying the stabilization of higher order switched systems by projecting the trajectory of the switched system.

Furthermore, a counter example is presented in order to prove that the main result cannot be extended to the case where the family of matrices $\{B_p : p \in \mathcal{P}\}$ are non-stable.

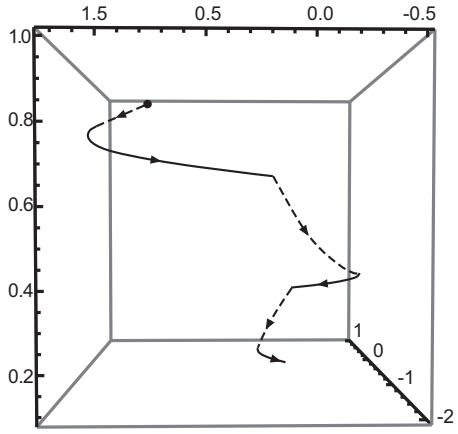


Figure 2: Solution of the switched system in Example 2.

References

[1] D. Angeli, D. Liberzon, *A note on uniform global asymptotic stability of nonlinear switched systems in triangular form*, Proc. of the 14th Int. Symp. on Mathematical Theory and Networks and Systems (MTNS) (2000).

[2] S. Cong, *Characterising the stabilisability for second-order linear switched systems*, International Journal of Control, **86** (2013) 519–528.

[3] B. Hu, X. Xu, P. Antsaklis, A. Michel, *Robust stabilizing control laws for a class of second-order switched systems*, Systems Control Lett. **38** (1999) 197–207.

[4] Z. G. Li, C. Y. Wen, Y. C. Soh, *Stabilization of a class of switched systems via designing switching laws*, IEEE Trans. Automat. Contr. **46** (2001) 665–670.

[5] D. Liberzon, J. P. Hespanha, A. S. Morse, *Stability of switched linear systems: a Lie-algebraic condition*, Systems Control Lett. **37** (1999) 117–122.

[6] D. Liberzon, A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Syst. Mag. **19** (1999) 59–70.

[7] D. Liberzon, *Switching in systems and Control*, Birkhauser, Boston 2003.

[8] H. Lin, P. J. Antsaklis, *Stability and stabilizability of switched linear systems: a survey of recent results*, IEEE Transactions on Automatic Control **54** (2009) 308–322.

[9] C. Pérez, F. Benítez, *Switched convergence of second-order switched nonlinear systems*, International Journal of Control, Automation, and Systems **10** (2012) 920–930.

[10] S. Sastry, *Nonlinear systems: Analysis, Stability, and Control*, Springer-Verlag, New York 1998.

[11] Z. Sun, S. S. Ge, *Switched linear systems*, Springer 2005.

[12] Z. Sun, S. S. Ge, *Stability theory of switched dynamical systems*, Springer 2011.

[13] M. A., Wicks, P. Peleties, R. A. DeCarlo, *Construction of piecewise Lyapunov functions for stabilizing switched systems*, Proc. of the 33rd Conf. on Decision and Control (1994) 3492–3497.

[14] M. A. Wicks, R. A. DeCarlo, *Solution of coupled Lyapunov equations for the stabilization of multimodal linear systems*, Proc. of 1997 American Control Conf. (1997) 1709–1713.

[15] X. Xu, P. Antsaklis, *Stabilization of second-order LTI switched systems*, International Journal of Control (2000) 1261–1279.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An energy evaluation of data-parallel applications in heterogeneous systems

Borja Pérez[1], Esteban Stafford[1], Jose Luis Bosque[1] and Ramón Beivide[1]

[1] *Computer and Electronics Engineering Department, Universidad de Cantabria*

emails: perezpavonb@unican.es, stafforde@unican.es, bosquejl@unican.es, beividej@unican.es

Abstract

The use of heterogeneous systems is on the rise as they improve both performance and energy efficiency. However, the programming of these machines requires considerable effort to get the best results in massively data-parallel applications. Maat is a library that enables OpenCL programmers to efficiently execute single data-parallel kernels using all the available devices on a heterogeneous system. It provides the programmer with an abstract view of the system to manage both multi-platform and heterogeneous environments, regardless of the underlying architecture. In addition, the library offers a set of load balancing methods, which perform the data partitioning and distribution among the devices. These allow utilizing more of the system's performance and consequently reduce execution time. This paper studies whether the use of all the devices in the heterogeneous system to execute a single kernel improves the energy efficiency, as it does the performance.

Key words: Heterogeneous systems, load balancing, energy efficiency, OpenCL.

1 Introduction

Nowadays, heterogeneous systems are commonly built by combining the computing power of general purpose CPUs with hardware accelerators (usually, GPUs) [1]. This architecture provides very high performance thanks to the large number of cores available in the GPU coupled with a very low energy consumption, which maximizes energy efficiency [2]. However, this computing power can only be exploited through highly parallel workloads. Hence the need of general purpose processors.

Harnessing the computing power of these systems is a considerable challenge by itself. The most common programming model, being used by CUDA [3] and OpenCL [4], is the host-device model. It dictates that the host processor starts the execution of an application, and offloads highly parallel parts to the GPU. In general, the host waits for the completion of these code sections. Consequently, this model does not exploit the computing power of the host, as it remains idle while the GPU is in operation. Interestingly, the CPUs are still consuming a noticeable amount of power during these periods, significantly reducing the energy efficiency of the system as a whole.

Maat is an OpenCL library that hides from the developers all the different computing elements of a heterogeneous system, which guarantees code portability and improves productivity [5]. With Maat, massively data-parallel applications can then be programmed using the host-device model, while it conveniently distributes the workload among all the available computing elements. The distribution is performed by a load balancing algorithm that can be configured by the programmer, much like OpenMP. Maat offers a set of load balancing algorithms that can be used for applications of different nature. Choosing and adequately configuring the algorithm is fairly simple and can considerably improve the performance of the system. This is because all devices perform useful work all the time, and no device remains idle waiting for the completion of another.

However, using all devices simultaneously to solve a single problem increases the energy consumption considerably. To better understand the implications of different load-balancing strategies, this work provides a study of the energy efficiency of the whole heterogeneous system, composed by several CPUs and GPUs. In order to widen the study, the GPUs are operated at different frequencies, thus varying their apparent performance. To measure the energy efficiency, this study measures the execution time of several benchmarks, as well as the energy consumed by the system. The execution of each benchmark is repeated with different load balancing algorithms and GPU frequencies. These results are compared with the baseline scenario that consists of a heterogeneous system in which only one GPU is used.

The experimental results show some interesting conclusions. The first is that for all the benchmarks and all the frequencies analysed, there is a load balancing algorithm that improves the baseline results, both in execution time and energy consumption. Second, the higher frequency on the GPU, greatly reduces the total energy consumption for all benchmarks analysed. The higher power consumption throughout the execution is compensated by reduction in execution time, thereby reducing the total energy and thus improving energy efficiency.

The rest of the paper is organised as follows. A description of the different load balancing methods implemented in Maat is presented in Section 2. Followed by Section 3 that describes the experimental evaluation and analyses the most relevant results. Finally, Section 4 shows the most interesting conclusions of this study and proposes future work.

2 Load Balancing Algorithms

Extracting the best performance out of heterogeneous systems is conditioned by an adequate use of all their resources. The following considerations are vital to making such an optimal workload distribution:

- Devices may have significant performance and architectural differences, unknown at design time.
- Accurately modelling the performance of the different devices is particularly difficult if the workload is irregular.
- Data communication with the devices is performed over slow interconnects.

The first two issues encourage the use of *dynamic* algorithms, which distribute the workload to the devices during runtime. Dynamic algorithms do not require prior information about the performance of the devices. However, the last consideration can impose a significant overhead. For regular workloads, the first two issues are less relevant so a *static* algorithm is worth considering. This can give better results than the dynamic due to a minimisation of the communication overhead. Trying to balance the above extremes, a third algorithm, named *h-guided*, has been implemented.

All three algorithms focus on the balancing of data parallel workloads, in which every device performs the same computation on a disjoint partition of the data. Thus, data are themselves the subjects of the distribution.

2.1 Static Balancing

This algorithm is based on dividing the load in as many portions as devices are available in the system. Then, it assigns a single portion to each of them. In order to obtain good performance, the load has to be divided in such a way as to make every device finish their computation simultaneously. Otherwise some devices will be idle while the computation is being completed. This is achieved by assigning each device a portion of work proportional to its computational power.

Let there be a heterogeneous system $H = \{D, P\}$, where $D = \{d_1, \cdots d_n\}$ is the set of devices and $P = \{P_1, \cdots P_n\}$ are the corresponding *computational powers*. P_i is defined as the amount of work that each device can complete in a time unit, including the necessary communication overhead. For a given application, these depend on the architecture of the devices, and are parameters that must be given to the model.

In order to perform the data partition, consider a given work-load, that needs to process W *work-items* grouped in G *work-groups*. In OpenCL, each group can be executed concurrently as they do not require communication among them. Whereas threads within

 ISBN: 978-84-608-6082-2

a group can synchronize among themselves. Consequently a work-group should not be split across devices, and it should be considered as the atomic distribution unit. All groups have the same size $L_s = \frac{W}{G}$, called *local work size*.

The response time of the heterogeneous system will be that of the last device to finish its work, $T_H = max_{i=1}^{N} T_{d_i}$. Therefore, the goal of the static method is to obtain a mapping of work-groups to devices, so that the workload is best balanced. This means, to find a tuple $\{\alpha_1, \cdots \alpha_n\}$, where α_i is the number of work-groups assigned to the device d_i, such that all the devices finish their work at the same time, and then the system response time is minimized:

$$T_H = T_{d_1} = \cdots = T_{d_n} \Rightarrow \frac{L_s \cdot \alpha_1}{P_1} = \cdots = \frac{L_s \cdot \alpha_n}{P_n}$$

Following the optimal algorithm proposed by O. Beaumont in [6] this can be done with complexity $O(n^2)$ in two steps:

- First, α_i is calculated so that $\frac{\alpha_i}{P_i}$ is almost constant $\forall\ i \in [1, \cdots n]$, and $\alpha_1 + \alpha_2 + \cdots + \alpha_n \leq G$:

$$\alpha_i = \left\lfloor \frac{P_i}{\sum_{i=1}^{n} P_i} \cdot G \right\rfloor$$

- Second, if $\sum_{i=1}^{n} \alpha_i < G$, then the remaining work-groups are assigned to the most powerful device. This amount of work is practically negligible, and does not disturb the load distribution.

This algorithm guarantees that the number of synchronization points is minimized, and performs well when facing regular loads, provided computational power of the devices are accurately known. However, it is not adaptable, so performance is not so good for irregular loads.

2.2 Dynamic Balancing

This algorithm divides the load in small, equally-sized packages, many more than the amount of available devices. The runtime orchestration is carried out by a *master thread* that follows the next algorithm:

1. The master splits the number of work-groups G, in a set of p packages, all of them with the same size $Package\_size = \left\lfloor \frac{G}{p} \right\rfloor$. If G is not divisible by p, an extra package will have the remainder of the division.

2. The master launches one package on each device.

3. The master waits for the completion of any package.
4. When device d_i completes the execution of a package:
   (a) The master stores results returned by the device.
   (b) If there are outstanding packages the next package is assigned to device d_i.
   (c) Else, if d_i is a GPU and there is a busy CPU d_i steals the package from the CPU.
   (d) If none of this conditions is met, the master proceeds to step 5.
   (e) The master returns to step 3.
5. The master ends as all the packages have been processed and their results are stored in the host.

This shows that the dynamic approach can adapt to different hardware and workload scenarios where the static can not. Attending to performance, the communication overhead must be taken into account. In the static, there is one package per device, but in the dynamic there are many more. Even if the data volume transferred to and from the devices is the same, the dynamic has a greater time dedicated to synchronization than the static.

2.3 H-guided Balancing

The h-guided algorithm strives to reduce the amount of synchronization points inherent to the dynamic scheme. It can be thought of as a refinement of the latter, as it revolves about the same basic algorithm, only there is a difference in the size of the packages. These are not equal, but of diminishing size. On a first approach they can be computed as $Package\_size = \left\lfloor \frac{G_r}{n} \right\rfloor$. Where G_r is the number of pending work-groups and n is the number of available devices. This way, the packages are moderately big at the beginning of the execution and small at the end. This results in a reduction in synchronization points, while maintaining adaptability mostly at the end of the execution, when a finer-grained load distribution is needed. The size diminishes down to a minimum size that the algorithm must be provided with.

This solution has been used already in homogeneous systems, but when applied to heterogeneous machines it needs a further refinement. If there is a great difference in the computational power of the devices, a big package may be assigned to a slow device, delaying the completion of the whole program. To avoid this, the h-guided algorithm takes into account the computational power of the device, in a similar fashion to the static approach. Then, considering P_i as the computational power of device d_i, the size of the packages is calculated as:

$$Package\_size = \left\lfloor \frac{P_i}{\sum_{i=1}^{n} P_i} \cdot \frac{G_r}{n} \right\rfloor$$

 ISBN: 978-84-608-6082-2

3 Performance and Energy Evaluation

3.1 Experimental Set-up

The experiments presented in this section have been performed on a system with two GPUs, two CPUs and 16 GBs of DDR3 memory. The CPUs are Intel Xeon E5-2620, with six cores that can run two threads at 2.0 GHz. The CPUs are connected via QPI, which results in OpenCL detecting them as a single device. Therefore, throughout the remainder of this document, any reference to the CPU includes both Xeon E5-2620 processors. The GPUs are two NVIDIA K20m with 13 SIMD lanes and 5 GBytes of VRAM each. The GPUs are connected to the system using independent PCI 2.0 slots.

Four applications have been chosen for the experiments. Two of them are part of the AMD APP SDK[7]. Both, NBody and MatMul, are well-known and regular applications in which different, equal-sized work units have the same running times. The other two applications, which are in-house implementations of known algorithms, are examples of irregular workloads in which different, equal-sized work units may have different running times. First, RAP is an implementation of the Resource Allocation Problem, based on the one proposed by Acosta *et al.* [8]. It must be noted that there is a certain pattern in the irregularity of RAP, as each successive package represents a bigger amount of work than the previous. Finally, a raytracing algorithm (RAY) was implemeted as an example of a truly irregular workload. This computes a realistic rendering of a scene by following light rays with independent threads. Thus, each of them represents an unpredictable amount of work, as the number of ray bounces depends on the objects of the scene.

The performance has been measured as the *speedup* with respect to the baseline execution time, using OpenCL and a single GPU. Note that the maximum speedup of the system is not the number of devices, due to different computational power of each of them. The CPUs have significantly less performance than the GPUs for the considered loads, and the performance difference is application-dependent.

To measure the energy consumption, a monitor was developed that samples the power consumption of each device. This periodically measures the GPU power sensors through the NVIDIA Management Library (NVML) [9]. It also reads the Running Average Power Limit (RAPL) registers of the CPUs [10]. The monitor is able to restrict the measurements of the energy consumed to a given *Region Of Interest (ROI)*. Finally, the energy efficiency is represented in terms of Energy-Delay-Product (EDP) [11].

All the experiments have been performed with two different GPU frequencies, 324 and 758 MHz, that correspond with the minimum and maximum frequency supported by the GPU. At the highest frequency, the GPU has much more computing power than the CPU, so reducing the frequency makes the system less heterogeneous. That is, the computing capacity of both devices becomes more similar. Depending on the application, this can have a strong impact in performance.

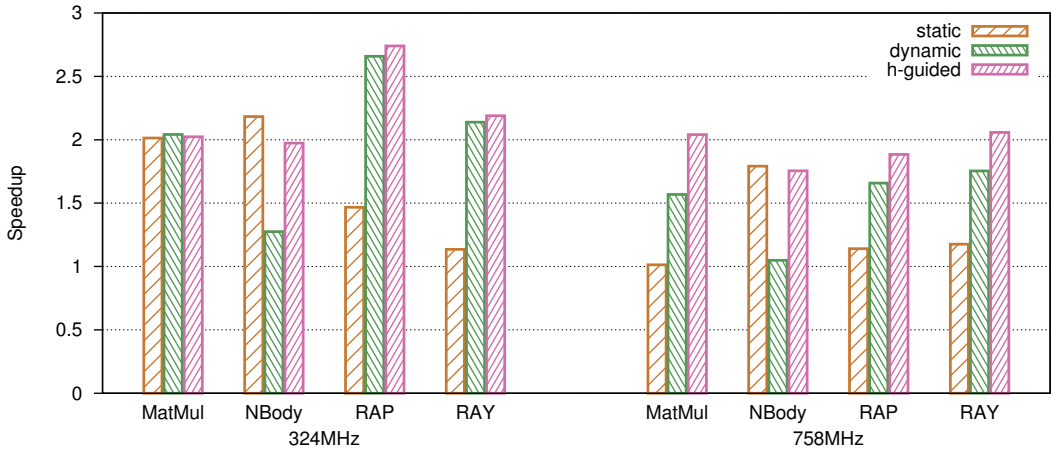


Figure 1: Speedup for each algorithm relative to a single GPU.

3.2 Experimental Results

Figure 1 presents the speedup of each load balancing algorithm and benchmark referred to the performance of a single GPU. These results are presented for the two different frequencies, 324 MHz and 758 MHz. It is important to highlight that for any of the analysed cases, there is always a load balancing algorithm that improves the performance of the baseline scenario.

The speedup of the regular applications is limited by both the sensitivity to the number of packages and the performance differences between devices, that is larger with 758 MHz frequency. Observe that the behaviour of NBody is very similar in both frequencies, and that the best algorithm is the static, although closely followed by the h-guided. However for MatMul, this behaviour depends strongly on the frequency. The poor performance shown for maximum frequency can be attributed to the enormous performance difference between GPUs and CPUs. In this case, even the smallest work package is too big for the CPU, and therefore it is better not to use the CPUs altogether. In the dynamic and h-guided methods the GPUs end up carrying out all the work, because they are able to steal packages initially scheduled to CPUs. On the other hand, when the low frequency of the GPUs is used and the computing power of the devices is comparable, the CPU can perform a significant portion of the work and contribute to the performance of ofthe system. In this case, the three algorithms present similar speedups.

The analysis of the irregular applications RAP and RAY is very similar for both frequencies, from which the h-guided method obtains the best results. This is because the size of the packages assigned is proportional to the computing power of the device and inversely proportional to the number of devices. As a consequence, more devices imply a

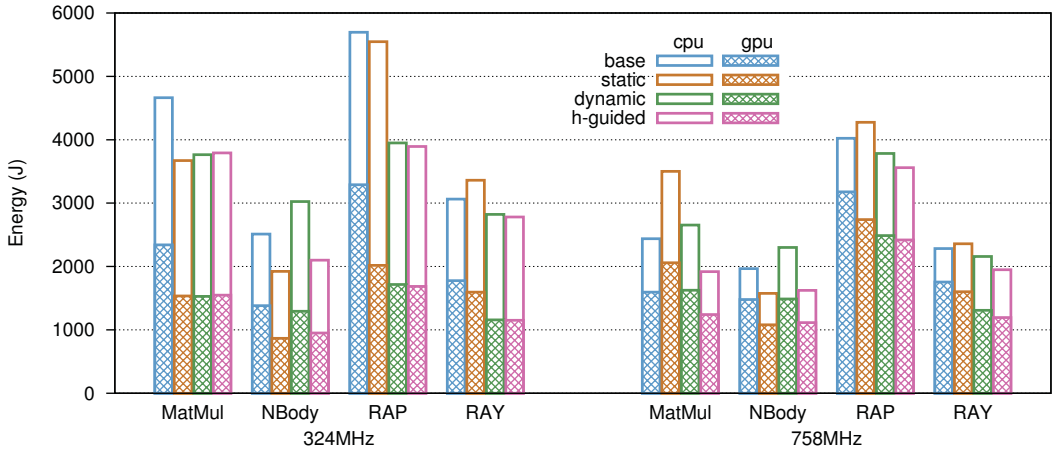


Figure 2: Energy consumption for each algorithm and benchmark.

greater amount of smaller packets, enabling a finer load balancing. Thus, the CPUs can contribute processing small packages, while the majority of the computation is carried out by the GPUs. The only difference between both figures is that the speedup of the dynamic is closer to the h-guided with the minimum frequency. This is because deceasing the GPU frequency reduces the differences in computational capabilities between both devices, and then reduces the heterogeneity of the system. This causes a reduction in the size difference of the packages in h-guided, thus reducing the impact in performance and the dynamic can reach a similar performance.

Attending to energy considerations, Figure 2 shows the energy consumption of the system for each benchmark, load balancing algorithm and GPU frequency (less is better). The bars on the graph are divided to separately show the energy corresponding to the twelve CPU cores and the two GPUs. The bar labeled as "base" corresponds to the results with one GPU.

The first conclusion to highlight is that, despite using two GPUs, the total energy consumption of the heterogeneous system is less than that of the "base". Therefore, there is at least one load balancing algorithm that improves the "base" energy consumption. This comes as a consequence of two improvements: the reduction in the execution time of the benchmarks, and that all the devices are contributing useful work, improving the energy efficiency of the system. This behaviour is exacerbated with the maximum GPU frequency, because it reduces considerably the execution time. However, the energy consumption is relatively less with the minimum frequency, as the contribution of the CPU is more significant, thus resulting in an execution time shorter than that of the "base" (see Figure 1).

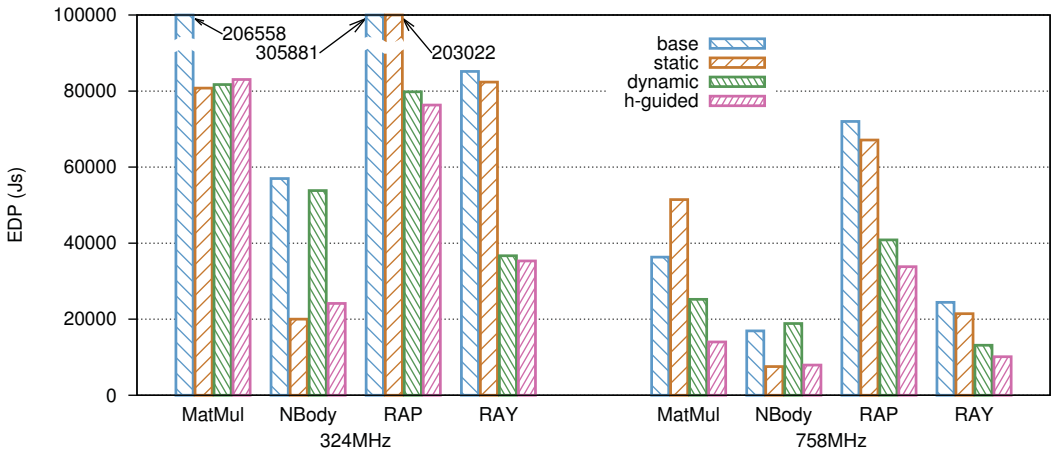


Figure 3: Energy Delay Product (EDP) for each algorithm and benchmark.

Energy saving is more noticeable in regular applications. In NBody the static algorithm reduces the base consumption in 30.5% and 24.7% for each frequency, while the h-guided reduces 19.6% and 21,1%. This is caused by the regularity of the applications, as the static algorithm can achieve an almost perfectly balanced distribution with minimal interaction between the devices. The h-guided saves up to 21.1% of the "base" energy in the MatMul benchmark for 758 MHz. But for 324 MHz the best result is given by the static with an improvement of 27%, yet closely followed by dynamic with 23.8% and the h-guided with 23%. Again, this is favoured by regularity of the application, but also by the small amount of packages necessary to get a good balance.

Regarding irregular benchmarks, RAP, despite having greater speedup than NBody yields a smaller energy reduction, around 13% for 758 MHz. This stems from the fact that the irregularity of the application requires to work with a large number of packages to obtain a good load balancing. This, in turn, causes an increase of the energy consumption with a greater interaction between CPU that does not perform useful work. However this small reduction increases dramatically for 324 MHz, reaching 46% with h-guided and 44% with the dynamic method. This behaviour can be explined again by the similarity of the computational capabilities of both devices, that also provides an important increment of the speedup (See Figure 1). For the ray-tracing benchmark, the situation is very similar but the greater reductions are obtained for 758 MHz, with 17.1% and 10%, respectively.

Finally, Figure 3 shows the results of the energy efficiency in terms of the Energy-Delay Product. This metric combines the performance and energy metrics in one value. Therefore this graphs shows a combination of the two previous ones, and most of the observations made above are confirmed by the results plotted in this graph.

In general, it can be said that there is always a load-balancing algorithm that obtains substantial EDP improvements compared to the "base" scenario. For regular applications the static algorithm is slightly better than the h-guided, except for the MatMul benchmark, whose behaviour has been explained above. And for irregular applications the h-guided algorithm reaches the best results for all the studied cases.

4 Conclusions

This paper presents an evaluation of the energy consumption of a set of load balancing algorithms for massively data-parallel applications in heterogeneous systems. By harnessing the computing power of the whole heterogeneous machine, these algorithms not only shorten the execution time of the applications, but also reduce their energy consumption, thus obtaining an improvement of the energy efficiency of the whole system.

The experimental results presented in this paper show that for both, regular and irregular applications, there is always a load balancing algorithm that reduces the execution time. This is a logical consequence of the full system having greater computing power than a single GPU. An interesting aspect shown in the paper is that a large performance difference between the devices increases the heterogeneity of the system, and therefore the h-guided has the best all-round performance.

Furthermore, the energy consumption of the machine with these algorithms is also reduced. These savings are more notable in regular applications, because the interaction between CPU and GPU to obtain a balanced workload is lower. Finally, energy efficiency results, expressed in terms of EDP, show that for all applications there is at least one load balancing algorithm that achieves substantial improvements.

The best overall results are obtained with the h-guided algorithm, yet the dynamic also gives very good results in irregular applications. Furthermore, dynamic does not require prior knowledge of the power of the computing devices, thus making it a good first approach to an unknown system. The static algorithm is appropriate for homogeneous environments and regular applications.

Future work will afford the study of more GPU frequencies as well as benchmarks in order to confirm the conclusions obtained in this paper. Additionally, new load balancing methods that take into account energy efficiency together with performance will be included in Maat library.

Acknowledgements

This work has been supported by the Spanish Science and Technology Commission under contract TIN2013-46957-C2-2-P, the University of Cantabria, grant CVE-2014-18166, the European Union FEDER (CAPAP-H5 network TIN2014-53522-REDT) and the European

HiPEAC Network of Excellence. The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under the Mont-Blanc Project (http://www.montblanc-project.eu), grant agreement n° 610402.

References

[1] J. Hestness, S. W. Keckler, and D. A. Wood. Gpu computing pipeline inefficiencies and optimization opportunities in heterogeneous cpu-gpu processors. In *Workload Characterization (IISWC), 2015 IEEE International Symposium on*, pages 87–97, Oct 2015.

[2] K. Ma, X. Li, W. Chen, C. Zhang, and X. Wang. Greengpu: A holistic approach to energy efficiency in gpu-cpu heterogeneous architectures. In *2012 41st International Conference on Parallel Processing*, pages 48–57, Sept 2012.

[3] David B. Kirk and Wen-mei W. Hwu. *Programming Massively Parallel Processors: A Hands-on Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2st edition, 2013.

[4] Benedict R. Gaster, Lee W. Howes, David R. Kaeli, Perhaad Mistry, and Dana Schaa. *Heterogeneous Computing with OpenCL - Revised OpenCL 1.2 Edition*. Morgan Kaufmann, 2013.

[5] Borja Pérez, José Luis Bosque, and Ramón Beivide. Simplifying programming and load balancing of data parallel applications on heterogeneous systems. In *Proc. of the 9th Workshop on General Purpose Processing using GPU, Barcelona, Spain, March 12*, pages 42–51, 2016.

[6] Olivier Beaumont, Vincent Boudet, Antoine Petitet, Fabrice Rastello, and Yves Robert. A proposal for a heterogeneous cluster ScaLAPACK (dense linear solvers). *IEEE Trans. Computers*, 50(10):1052–1070, 2001.

[7] Amd accelerated parallel processing software development kit v2.9. Last accesed November 2015.

[8] Alejandro Acosta, Robert Corujo, Vicente Blanco, and Francisco Almeida. Dynamic load balancing on heterogeneous multicore/multiGPU systems. In Waleed W. Smari and John P. McIntire, editors, *HPCS*, pages 467–476. IEEE, 2010.

[9] NVIDIA. NVIDIA Management Library (NVML). Last accesed April 2016.

[10] Efraim Rotem, Alon Naveh, Doron Rajwan, Avinash Ananthakrishnan, and Eli Weissmann. Power management architecture of the 2nd generation Intel Core microarchitecture, formerly codenamed Sandy Bridge. In *IEEE Int. HotChips Symp. on High-Perf. Chips (HotChips˜2011)*, 2011.

[11] Emilio Castillo, Cristóbal Camarero, Ana Borrego, and Jose Luis Bosque. Financial applications on multi-cpu and multi-gpu architectures. *J. Supercomput.*, 71(2):729–739, February 2015.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The Generalized Hyers-Ulam Stability of Additive ρ-Functional Inequalities in Random Normed Spaces

Supak Phiangsungnoen[1] and Wiyada Kumam[2]

[1] *Department of Mathematics, Faculty of Liberal Arts, Rajamangala University of Technology Rattanakosin (RMUTR), 264 Chakkrawat Rd., Chakkrawat, Samphanthawong, Bangkok 10100, Thailand.*

[2] *Program in Applied Statistics, Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi, Thanyaburi, Pathumthani 12110, Thailand.*

emails: supak.pia@rmutr.ac.th, wiyada.kum@mail.rmutt.ac.th

Abstract

In this paper, we prove the generalized Hyers-Ulam stability the following additive ρ-functional inequalities:

$$\mu_{f(x+y)-f(x)-f(y)}(t) \geq \mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t)$$

and

$$\mu_{2f(\frac{x+y}{2})-f(x)-f(y)}(t) \geq \mu_{\rho(f(x+y)-f(x)-f(y))}(t)$$

where in random normed spaces by using fixed point method.

Key words: Cauchy-Jensen Additive Functional Inequality, Fixed Point, ρ-Functional Inequality, Random Norm Spaces, Stability
MSC 2000: 46T99, 47H10, 47H09, 54H25.

1 Introduction

In 1940, S.M. Ulam [1] proposed the following stability problem of functional equations:

Given a group G_1, a metric group G_2 with the metric $d(\cdot,\cdot)$ and a positive number ϵ, does there exist $\delta > 0$ such that, if a mapping $f : G_1 \to G_2$ satisfies $d(f(xy), f(x)f(y)) \leq \delta$

forall $x, y \in G_1$, then a homomorphism $h : G_1 \to G_2$ exists with $df(x), h(x) \leq \epsilon$ for all $x \in G_1$?

Letter on, Hyers [2] provided a partial solution to Ulam's problem for the case of approximately additive mappings in which G_1 and G_2 are Banach spaces. The result of Hyers was extended by Aoki [3], by considering the unbounded Cauchy difference for additive mappings and by Rassias [4] for linear mappings by considering the unbounded Cauchy difference. The paper work of Rassias [4] has had a lot of influence in the development of what we call the *generalized Hyers-Ulam stability* or *generalized Hyers-Ulam-Rassias stability* of functional equations.

The most famous functional equation is the Cauchy functional equation:

$$f(x+y) = f(x) + f(y),$$

whose solution are called *additive mapping*. The functional equation

$$f(\frac{x+y}{2}) = \frac{1}{2}f(x) + \frac{1}{2}f(y),$$

is called the Jensen equation. The stability problems of several functional equations have been extensively investigated by a number of authors and there are many interesting results concerning this problem.

In 2001, Gilányi [5] showed that if f satisfies the functional inequality

$$\|2f(x) + 2f(y) - f(x-y)\| \leq \|f(x+y)\|$$

then f satisfies the Jordan-von Neumann functional equation

$$2f(x) + 2f(y) = f(x+y) + f(x-y).$$

Afterwards there are several papers proved the Hyers-Ulam stability of the Cauchy additive functional inequality and Cauchy-Jensen additive functional inequality in different spaces [7–10]

The notion of random normed space goes back to Sherstnev [11] as well as the works published in [12–14] who were dulled from Menger [15], Schweizer and Sklar [12] works. After the pioneering works by several mathematicians including authors [16–21] who focused at probabilistic functional analysis, Alsina [17] considered the stability of a functional equation in probabilistic normed spaces and, in 2008, Mihet and Radu [22] considered the stability of a Cauchy additive functional equation in random normed space via fixed point method.

Fixed point alternative method is the one play an important to proved stability problems. By using fixed point alternative methods, the stability problems of several functional equations have been extensively investigated by a number of authors (see [23–25] and reference there in).

Let Δ^+ denote the space of all distribution functions, that is, the space of all mappings $F : \mathbb{R} \cup \{-\infty, +\infty\} \to [0, 1]$ such that F is left-continuous, non-decreasing on $\mathbb{R}$, $F(0) = 0$ and $F(+\infty) = 1$. Let D^+ is a subset of Δ^+ consisting of all functions $F \in \Delta^+$ for which $l^- F(+\infty) = 1$ where $l^- f(x)$ denotes the left limit of the function f at the point x, that is, $l^- f(x) = \lim_{t\to x^-} f(t)$. The space Δ^+ is partially ordered by the usual point- wise ordering of functions, that is, $F \leq G$ if and only if $F(t) \leq G(t)$ for all $t \in \mathbb{R}$. The maximal element for Δ^+ in this order is the distribution function ε_0 given by

$$\varepsilon_0(t) = \begin{cases} 0, & \text{if } t \leq 0; \\ 1 & \text{if } t > 0. \end{cases}$$

In this paper, we prove the generalized Hyers-Ulam stability the following additive ρ-functional inequalities:

$$\mu_{f(x+y)-f(x)-f(y)}(t) \geq \mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t) \tag{1.1}$$

$$\mu_{2f(\frac{x+y}{2})-f(x)-f(y)}(t) \geq \mu_{\rho(f(x+y)-f(x)-f(y))}(t) \tag{1.2}$$

in random norm spaces via fixed point approach.

2 Preliminaries

Now we give some the following notations about triangular norm (shortly, t-norm) and random normed space (RN-space) will be used in the sequel.

Definition 2.1 ([26]) *A mapping $T : [0, 1] \times [0, 1] \to [0, 1]$ is a continuous triangular norm (briefly, a t-norm) if T satisfies the following conditions:*

(T1) T is associative and commutative;

(T2) T is continuous;

(T3) $T(a, 1) = a$ for all $a \in [0, 1]$;

(T4) $T(a, b) \leq T(c, d)$ whenever $a \leq c$ and $b \leq d$ for all $a, b, c, d \in [0, 1]$.

Basic examples are Lukasievicz t-norm T_L, that is, $T_L(a, b) = \max\{a + b - 1, 0\}$ for all $a, b \in [0, 1]$ and the t-norms T_P, T_M, T_D defined as follows: $T_P(a, b) := ab$, $T_M(a, b) := \min\{a, b\}$

$$T_D(a, b) := \begin{cases} \min\{a, b\}, & \text{if } \max\{a, b\} = 1; \\ 0 & \textit{otherwise}. \end{cases}$$

 ISBN: 978-84-608-6082-2

Definition 2.2 *A random normed space (briefly, RN-space) is a triple (X, μ, T), where X is a linear space, T is a continuous t-norm, and μ is a mapping from X into D^+ such that, the following conditions hold:*

(R1) $\mu_x(t) = \varepsilon_0(t)$ *for all* $t > 0$ *if and only if* $x = 0$*;*

(R2) $\mu_{\alpha x}(t) = \mu_x(\frac{t}{|\alpha|})$ *for all* $x \in X$ *and* $\alpha \neq 0$*;*

(R3) $\mu_{x+y}(t+s) \geq T(\mu_x(t), \mu_y(s))$ *for all* $x, y \in X$ *and* $t, s \geq 0$.

Definition 2.3 *Let (X, μ, T) be an RN-space.*

(1) A sequence $\{x_n\}$ in X is said to be convergent *to x in X if, for every $\epsilon > 0$ and $\lambda > 0$, there exists a positive integer N such that $\mu_{x_n - x}(\epsilon) > 1 - \lambda$ where $n \geq N$.*

(2) A sequence $\{x_n\}$ in X is said to be Cauchy sequence *if, for every $\epsilon > 0$ and $\lambda > 0$, there exists a positive integer N such that $\mu_{x_n - x_m}(\epsilon) > 1 - \lambda$ where $n \geq m \geq N$.*

(3) An RN-space (X, μ, T) is said to be complete *if every Cauchy sequence in X is convergent to a point in X.*

Theorem 2.4 *[26] If (X, μ, T) is an RN-space and $\{x_n\}$ is a sequence such that $x_n \to x$, then $\lim_{n\to\infty} \mu_{x_n}(t) = \mu_x(t)$ almost everywhere.*

Theorem 2.5 *[27] Let (X, μ, T) is an RN-space such that every Cauchy sequence in X has convergent subsequence. Then (X, μ, T) is complete.*

In 1958, Luxemburg [28], introduced the concept of generalized metric space as following definition:

Definition 2.6 *Let X be a set. A function $d : X \times X \to [0, \infty]$ is called a generalized metric on X if d satisfies*

(i) $d(x, y) = 0$ *if and only if* $x = y$*;*

(ii) $d(x, y) = d(y, x)$ *for all* $x, y \in X$*;*

(iii) $d(x, z) \leq d(x, y) + d(y, z)$ *for all* $x, y, z \in X$.

The following fixed point theorem which was proved by Diaz and Margolis [29] and Jung[30] will play an important role.

Theorem 2.7 *(The alternative of fixed point [29, 30]) Let (X, d) be a complete generalized metric space and let $A : X \to X$ be a strictly contractive mapping with Lipschitz constant $L \in (0, 1)$ such that*

$$d(x_0, A(x_0)) < +\infty$$

for some $x_0 \in X$. Then we have the following:

(a) A has a unique fixed point in the set $Y := \{y \in X : d(x_0, y) < \infty\}$;

(b) for all $x \in Y$, the sequence $\{A^n(x)\}$ converges to the fixed point x^*;

(c) $d(x_0, A(x_0)) \leq \delta$ implies $d(x^*, x_0) \leq \frac{\delta}{1-L}$.

3 The Hyer-Ulam stability of additive ρ-functional inequalities.

In this section, using fixed point method, we prove the Hyer-Ulam stability of additive ρ-functional inequalities (1.1) and (1.2) in random normed spaces. Let ρ be a real number with $|\rho| < 1$. We need the following lemma to prove the main results.

Lemma 3.1 *Let X be a linear space and (Y, μ, T) be an RN-space. If a mapping $f : X \to Y$ satisfies the functional inequalities:*

$$\mu_{f(x+y)-f(x)-f(y)}(t) \geq \mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t), \tag{3.1}$$

for all $x, y \in X$ and all $t > 0$, where ρ be a real number with $|\rho| < 1$. Then f is Cauchy additive.

Proof 1 *Assume that $f : X \to Y$ satisfies (3.1). Letting $x = y = 0$ in (3.1), we get $\mu_{f(0)}(t) = \mu_0(t) = 1$. So, we have $f(0) = 0$. Letting $x = y$ in (3.1), we get*

$$\mu_{f(2x)-2f(x)}(t) \geq \mu_0(t) = 1,$$

So, we have $f(2x) - 2f(x) = 0$, for all $x, y \in X$ and $t > 0$. Replacing $x = \frac{x}{2}$, thus we have

$$f(\frac{x}{2}) = \frac{1}{2}f(x), \tag{3.2}$$

for all $x \in X$. It follows from (3.1) and (3.2), we get

$$\begin{aligned} \mu_{f(x+y)-f(x)-f(y)}(t) &\geq \mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t) \\ &= \mu_{\rho(f(x+y)-f(x)-f(y)}(t) \\ &= \mu_{f(x+y)-f(x)-f(y}(\frac{t}{|\rho|}) \end{aligned}$$

 ISBN: 978-84-608-6082-2

for all $t > 0$. By condition of D^+ and (R1), we have

$$\mu_{f(x+y)-f(x)-f(y)}(t) = 1$$

for all $t > 0$. So, we get $f(x + y) - f(x) - f(y) = 0$ and $f(x + y) = f(x) + f(y)$ for all $x, y \in X$.

Now, we prove the Hyers-Ulam stability of the additive ρ-functional inequality (3.1) in random normed spaces.

Theorem 3.2 *Let $\psi : X \times X \to D^+$ ($\psi(x, y)$ is denoted by $\psi_{x,y}$) such that there exists $\alpha < 1$*

$$\psi_{x,y}(\alpha t) \geq \psi_{\frac{x}{2},\frac{y}{2}}(t) \tag{3.3}$$

for all $x, y \in X$ and $t > 0$. Let $f : X \to Y$ be a mapping satisfying

$$\mu_{f(x+y)-f(x)-f(y)}(t) \geq \min\left\{\mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t), \psi_{x,y}(t)\right\} \tag{3.4}$$

for all $x, y \in X$ and $t > 0$. Then, $\mathcal{A}(x) := \lim_{n\to\infty} \frac{1}{2^n} f(2^n x)$ exists for all $x \in X$ and $\mathcal{A} : X \to Y$ is an additive mapping such that

$$\mu_{f(x)-\mathcal{A}(x)}(t) \geq \psi_{x,x}((2 - 2\alpha)t) \tag{3.5}$$

for all $x \in X$ and $t > 0$.

Proof 2 *Letting $y = x$ in (3.4), we get*

$$\mu_{\frac{f(2x)}{2}-f(x)}(\frac{t}{2}) \geq \psi_{x,x}(t) \tag{3.6}$$

for all $x \in X$ and $t > 0$. Consider the set $S := \{g : X \to Y\}$ and introduce the generalized metric on S:

$$d(g, h) = \inf\left\{k \in \mathbb{R}_+ : \mu_{g(x)-h(x)}(kt) \geq \psi_{x,x}(t)\right\}$$

for all $x \in X$ and $t > 0$, where, as usual, $\inf \emptyset = +\infty$. It is easy to show that (S, d) is complete.

Now, we consider the linear mapping $J : S \to S$ such that

$$Jg(x) = \frac{1}{2} g(2x)$$

for all $x \in X$. Now, we show that J is a strictly contractive self-mapping of S with the Lipschitzconstant $L = \frac{\alpha}{2}$. Indeed, let $g, h \in S$ be the mappings such that $d(g, h) = \epsilon$. Then we have

$$\mu_{g(x)-h(x)}(\epsilon t) \geq \psi_{x,x}(t)$$

or all $x \in X$ *and* $t > 0$. *Hence, we have*

$$\begin{aligned} \mu_{Jg(x)-Jh(x)}(\frac{\alpha}{2}\epsilon t) &= \mu_{\frac{1}{2}g(2x)-\frac{1}{2}h(2x)}(\frac{\alpha}{2}\epsilon t) \\ &= \mu_{g(2x)-h(2x)}(\alpha\epsilon t) \\ &\geq \psi_{2x,2x}(\alpha t) \geq \psi_{x,x}(t) \end{aligned}$$

for all $x \in X$ *and* $t > 0$. *That is* $d(g,h) = \epsilon \Rightarrow d(Jg, Jh) < \frac{\alpha}{2}d(g,h)$, *for all* $g, h \in S$. *It follows from (3.6) that*

$$\mu_{\frac{f(2x)}{2}-f(x)}(\frac{t}{2}) \geq \psi_{x,x}(t)$$

for all $x \in X$ *and* $t > 0$. *So,* $d(f, Jf) \leq \frac{1}{2}$.

There exists a mapping $\mathcal{A} : X \to Y$ *satisfying the following:*

(1) $\mathcal{A}$ *is a fixed point of* J, *that is,*

$$\mathcal{A}(2x) = 2\mathcal{A}(x) \tag{3.7}$$

for all $x \in X$. *The mapping* $\mathcal{A}$ *is a unique fixed point of* J *in the set*

$$Q = \{g \in S : d(f,g) < \infty\}.$$

This implies that $\mathcal{A}$ *is a unique mapping satisfying (3.7) such that there exists* $u \in (0, \infty)$ *satisfying*

$$\mu_{f(x)-\mathcal{A}(x)}(ut) \geq \psi_{x,x}(t)$$

for all $x \in X$ *and* $t > 0$;

(2) $d(J^n f, \mathcal{A}) \to 0$ *as* $n \to \infty$. *This implies the equality*

$$\lim_{n\to\infty} \frac{1}{2^n} f(2^n x) = \mathcal{A}(x)$$

for all $x \in X$;

(3) $d(f, \mathcal{A}) \leq \frac{1}{1-\alpha}d(f, Jf)$, *which implies the inequality*

$$d(g, \mathcal{A}) \leq \frac{1}{2-2\alpha},$$

and so it follows that

$$\mu_{f(x)-\mathcal{A}(x)}(\frac{t}{2-2\alpha}) \geq \psi_{x,x}(t) \tag{3.8}$$

 ISBN: 978-84-608-6082-2

which implies that the inequality (3.5) holds. By (3.4), we have

$$\begin{aligned}
&\mu_{\mathcal{A}(x+y)-\mathcal{A}(x)-\mathcal{A}(y)}\\
&= \lim_{n\to\infty} \mu_{\frac{1}{2^n}f(2^n(x+y))-\frac{1}{2^n}f(2^nx)-\frac{1}{2^n}f(2^ny)}(t)\\
&\geq \lim_{n\to\infty} \min\left\{\mu_{\rho(2f(2^{n-1}(x+y))-f(2^nx)-f(2^ny))}(t), \psi_{2^nx,2^ny}(\frac{t}{2^n})\right\}\\
&\geq \lim_{n\to\infty} \min\left\{\mu_{\rho(2f(2^{n-1}(x+y))-f(2^nx)-f(2^ny))}(t), \psi_{x,y}(\frac{t}{(2\alpha)^n})\right\}
\end{aligned}$$

for all $x, y \in X$, $t > 0$ and $n \geq 1$. Since $0 < \alpha < 1$, we have $\lim_{n\to\infty} \psi_{x,y}(\frac{t}{(2\alpha^n)}) = 1$. Hence

$$\mu_{\mathcal{A}(x+y)-\mathcal{A}(x)-\mathcal{A}(y)}(t) \geq \mu_{\rho(2\mathcal{A}(\frac{x+y}{2})-\mathcal{A}(x)-\mathcal{A}(y))}$$

for all $x, y \in X$ and $t > 0$. Therefore $\mathcal{A} : X \to Y$ is additive. This completes the proof.

Corollary 3.3 *Let $\phi : X \times X \to D^+$ ($\phi(x, y)$ is denoted by $\phi_{x,y}$) such that there exists $\alpha < 1$*

$$\phi_{x,y}(\alpha t) \geq \phi_{\frac{x}{2},\frac{y}{2}}(t) \tag{3.9}$$

for all $x, y \in X$ and $t > 0$. Let $f : X \to Y$ be a mapping satisfying

$$\mu_{f(x+y)-f(x)-f(y)}(t) \geq \min\left\{\mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t), \frac{t}{t+\phi_{x,y}(t)}\right\} \tag{3.10}$$

for all $x, y \in X$ and $t > 0$. Then, $\mathcal{A}(x) := \lim_{n\to\infty} \frac{1}{2^n} f(2^nx)$ exists for all $x \in X$ and $\mathcal{A} : X \to Y$ is an additive mapping such that

$$\mu_{f(x)-\mathcal{A}(x)}(t) \geq \frac{(2-2\alpha)t}{(2-2\alpha)t + \phi_{x,x}(\alpha t)} \tag{3.11}$$

for all $x \in X$ and $t > 0$.

Proof 3 *The prove follows from Theorem 3.2 by replacing $\frac{t}{t+\phi_{x,y}(t)}$ instead $\psi_{x,y}(t)$ for all $x, y \in X$ and $t > 0$.*

Corollary 3.4 *Let $\theta \geq 0$ and let p be a real number with $p > 1$. Let X be a normed vector space with the norm $\|\cdot\|$. Let $f : X \to Y$ be an odd mapping satisfying*

$$\mu_{f(x+y)-f(x)-f(y)}(t) \geq \min\left\{\mu_{\rho(2f(\frac{x+y}{2})-f(x)-f(y))}(t), \frac{t}{t+\theta(\|x\|^p+\|y\|^p)}\right\} \tag{3.12}$$

for all $x, y \in X$ and $t > 0$. Then, $\mathcal{A}(x) := \lim_{n\to\infty} \frac{1}{2^n} f(2^nx)$ exists for all $x \in X$ and $\mathcal{A} : X \to Y$ is an additive mapping such that

$$\mu_{f(x)-\mathcal{A}(x)}(t) \geq \frac{(2-2^p)t}{(2-2^p)t + 2\theta\|x\|^p} \tag{3.13}$$

for all $x \in X$ and $t > 0$.

Proof 4 *The prove follows from Corollary 3.2 by tacking* $\phi_{x,y}(t) := \theta(\|x\|^p + \|y\|^p)$ *for all* $x, y \in X$ *and* $t > 0$.

Theorem 3.5 *Let* $\psi : X \times X \to D^+$ *(*$\psi(x, y)$ *is denoted by* $\psi_{x,y}$*) such that there exists* $\alpha < 1$

$$\psi_{x,y}(\alpha t) \geq \psi_{2x,2y}(t) \tag{3.14}$$

for all $x, y \in X$ *and* $t > 0$. *Let* $f : X \to Y$ *be a mapping satisfying (3.4). Then,* $\mathcal{A}(x) := \lim_{n\to\infty} 2^n f(\frac{x}{2^n})$ *exists for all* $x \in X$ *and* $\mathcal{A} : X \to Y$ *is an additive mapping such that*

$$\mu_{f(x)-\mathcal{A}(x)}(t) \geq \psi_{x,x}((\frac{2-2\alpha}{\alpha})t) \tag{3.15}$$

for all $x \in X$ *and* $t > 0$.

Proof 5 *We could prove the same statement with the same manner in spite of replacing the condition* $\psi_{x,y}(\alpha t) \geq \psi_{2x,2y}(t)$ *into* $\psi_{x,y}(\alpha t) \geq \psi_{\frac{x}{2},\frac{y}{2}}(t)$ *by defining* J *such that* $Jg(x) = \frac{1}{2}g(2x)$ *instead of* $Jg(x) = 2g(x)$. *It could be also applied to Theorem 3.2.*

Lemma 3.6 *Let* X *be a linear space and* (Y, μ, T) *be an RN-space. If a mapping* $f : X \to Y$ *satisfies* $f(0) = 0$ *and*

$$\mu_{2f(\frac{x+y}{2})-f(x)-f(y)}(t) \geq \mu_{\rho(f(x+y)-f(x)-f(y))}(t), \tag{3.16}$$

for all $x, y \in X$ *and all* $t > 0$. *Then* f *is Cauchy additive.*

Proof 6 *The prove follows from Lemma 3.1. Assume that* $f : X \to Y$ *satisfies (3.16). Letting* $y = 0$ *in (3.16), we get* $\mu_{2f(\frac{x}{2})-f(x)}(t) \geq \mu_0(t) = 1$. *So, we have*

$$f(\frac{x}{2}) = \frac{1}{2}f(x),$$

for all $x \in X$.

It follows from (3.1) and (3.16), so we have that $f(x+y) = f(x)+f(y)$ *for all* $x, y \in X$.

Theorem 3.7 *Let* $\psi : X \times X \to D^+$ *(*$\psi(x, y)$ *is denoted by* $\psi_{x,y}$*) such that there exists* $\alpha < 1$

$$\psi_{x,y}(\alpha t) \geq \psi_{2x,2y}(2t) \tag{3.17}$$

for all $x, y \in X$ *and* $t > 0$. *Let* $f : X \to Y$ *be a mapping satisfying*

$$\mu_{2f(\frac{x+y}{2})-f(x)-f(y)}(t) \geq \min\left\{\mu_{\rho(f(x+y)-f(x)-f(y))}(t), \psi_{x,y}(t)\right\} \tag{3.18}$$

for all $x, y \in X$ *and* $t > 0$. *Then,* $\mathcal{A}(x) := \lim_{n\to\infty} 2^n f(\frac{x}{2^n})$ *exists for all* $x \in X$ *and* $\mathcal{A} : X \to Y$ *is an additive mapping such that*

$$\mu_{f(x)-\mathcal{A}(x)}(t) \geq \psi_{x,x}((1-\alpha)t) \tag{3.19}$$

for all $x \in X$ *and* $t > 0$.

 ISBN: 978-84-608-6082-2

Proof 7 *Since f is odd mapping and $f(0) = 0$. Letting $y = 0$ in 3.18, we have*

$$\mu_{2f(\frac{x}{2})-f(x)}(t) \geq \mu_{\psi_{x,0}(t)}$$

for all $x \in X$.

Consider the set $S := \{g : X \to Y\}$ and introduce the generalized metric on S:

$$d(g,h) = \inf\left\{k \in \mathbb{R}_+ : \mu_{g(x)-h(x)}(kt) \geq \psi_{x,0}(t)\right\}$$

for all $x \in X$ and $t > 0$, where, as usual, $\inf \emptyset = +\infty$. It is easy to show that (S,d) is complete.

Now, we consider the linear mapping $J : S \to S$ such that

$$Jg(x) = 2g(\frac{x}{2})$$

for all $x \in X$.

Let $\mathcal{A}(x) : X \to Y$ be defined as in the proof of Theorem 3.2. Then $\mathcal{A}$ is Cauchy additive,as desired.

Acknowledgements

The author was supported by Rajamangala University of Technology Rattanakosin Research and Development Institute.

References

[1] S. M. Ulam, *Problems in Modern Mathematics*, John Wiley and Sons, New York, USA, 1964.

[2] D. H. Hyers, *On the stability of the linear functional equation*, Proceedings of the National Academy of Sciences of the United States of America, **27(4)**(1941), 222–224.

[3] T. Aoki, *On the stability of the linear functional equation*, Proc. Natl. Acad. Sci. USA, **27**(1941), 222–224.

[4] Th.M Rassias, *On the stability of the linear mapping in Banach spaces*. Proc Am Math Soc., **72**(1978), 297–300.

[5] A. Gilányi, *Eine zur Parallelogrammgleichung äquivalente Ungleichung*, Aequationes Math. **62** (2001), 303–309.

[6] W. Fechner, *Stability of a functional inequalities associated with the Jordan-von Neumann functional equation*, Aequationes Math. **71** (2006), 149–161.

[7] A. Gilányi, *On a problem by K. Nikodem*, Math. Inequal. Appl. **5** (2002), 707–710.

[8] C. Park, Y. Cho and M. Han, *Stability of functional inequalities associated with Jordan-von Neumann type additive functional equations*, J. Inequal. Appl. (2007), Art. ID 41820 (2007).

[9] J. Choi, J. Seong,C. Park, *Additive ρ-functional inequalities in normed spaces*, J. Nonlinear Sci. Appl. **9**(2016), 247–253.

[10] J-H. Kim, G.A. Anastassiou, C. Park, *Additive ρ-functional inequalities in fuzzy normed spaces*, J. Computational Analysis and Applications **21(6)**(2016), 1115–1126.

[11] A.N. Šerstnev, *On the motion of a random normed space*, Dokl. Akad. Nauk SSSR, **149**(1963), 280– 283.

[12] B. Schweizer, A. Sklar, *Probabilistic metric spaces*, Elsevier, North Holland, 1983.

[13] O.Hadžić, E.Pap, *Fixed point theory in PM-spaces*, Kluwer Academic, Dordrecht, 2001.

[14] O.Hadžić, E. Pap, *New classes of probabilistic contractions and applications to random operators*, Fixed Point Theory and Applications, ed. by Y.J. Cho, J.K. Kim, S.M. Kong, Nova Science Publishers, Hauppauge, **4**(2003), 97–119.

[15] K.Menger, *Statistical metrics*, Proc. Natl. Acad. Sci. USA, **28**(1942), 535–537.

[16] C.Alsina, *On the stability of a functional equation a rising in probabilistic normed spaces*, in General Inequalities, Oberwolfach, **(5)**1986, 263–271.

[17] C. Alsina, B. Schweizer, A. Sklar, *On the definition of a probabilistic normed space*, Aequ. Math., **46**(1993), 91–98.

[18] C. Alsina, B. Schweizer, A. Sklar, *Continuity properties of probabilistic norms*, J. Math. Anal. Appl., **208**(1997), 446–452.

[19] B. Lafuerza-Guillen, A. Rodriguez-Lallena, C.Sempi, *A study of boundedness in probabilistic normed spaces*, J. Math. Anal. Appl., **232**(1999), 183–196.

[20] B. Lafuerza-Guillén, *D-bounded sets in probabilistic normed spaces and their products*, Rend. Mat., Ser. VII, **21**(2001), 17–28.

[21] B.Lafuerza-Guillén, *Finite products of probabilistic normed spaces*, Rad.Mat., **13**(2004), 111–117.

 ISBN: 978-84-608-6082-2

[22] D. Mihet, V. Radu, *On the stability of the additive Cauchy functional equation in random normed spaces*, J Math Anal Appl., **343**(2008), 567–572.

[23] J. Brzdek, J.Chudziak, Z. Pales, *A fixed point approach to stability of functional equations*, Nonlinear Anal.(TMA), **74**(2011), 6728–6732.

[24] L. Cădariu, V. Radu, *Fixed points and the stability of the Cauchy functional equation : a fixed point approach*, Grazer Math, Berichte, **346**(2004), 43–52.

[25] V. Radu, *The fixed point alternative and the stability of functional equations*. Sem Fixed Point Theory., **4(1)**(2003), 91-96.

[26] B. Schweizer, A. Sklar, *Statistical metric spaces*, Pac J Math., **10**(1960), 313–334.

[27] Y.J. Cho, Th.M. Rassias and R. Saadati, *Stability of functional equations in random normed spaces*, Springer Optimization and Its Applications, **86**(2013).

[28] W.A.J. Luxemburg, *On the convergence of successive approximations in the theory of ordinary differential equations II*, Proc. K. Ned. Akad. Wet., Ser. A, Indeg. Math., **20**(1958), 540–546.

[29] J.B. Diaz, B. Margolis, *A fixed point theorem of the alternative for contractions on generalized complete metric space*, Bull Am Math Soc., **126(74)**(1968), 305–309.

[30] C.F.K. Jung, *On generalized complete metric spaces*, Bull. Am. Math. Soc., **75**(1969),113–116.

 ISBN: 978-84-608-6082-2

Volume IV

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Tile partition analysis for a parallel HEVC encoder

Pablo Piñol[1], Otoniel López-Granado[1], Héctor Migallón[1], Vicente Galiano[1] and Manuel P. Malumbres[1]

[1] *Department of Physics and Computer Architecture, Miguel Hernández University*

emails: pablop@umh.es, otoniel@umh.es, hmigallon@umh.es, vgaliano@umh.es, mels@umh.es

Abstract

The new video coding standard HEVC introduces a new concept named "tiles". Tiles are rectangular regions of a video frame which can be encoded in an independent way. In order to reduce the time needed to encode a video sequence with HEVC, we have used a parallel approach based on tiles, and have evaluated the benefits and drawbacks of this approach. In our tests we obtain speed ups of up to 9.3x using 10 processes with a low R/D loss.

Key words: tiles, HEVC, video coding, Parallel algorithms, multicore, performance

1 Introduction

The Joint Collaborative Team on Video Coding (JCT-VC) has developed a new video coding standard, named High Efficiency Video Coding (HEVC) [1]. JCT-VC is formed by members of the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG). The new standard achieves nearly a 50% of bit rate saving, when compared to the previous video coding standard H.264/AVC (Advanced Video Coding) [2]. The increase in the efficiency is bound to an increase in the computational complexity. To address the increase in complexity we make use of parallelization techniques, and thus, take advantage of the available parallel computing architectures.

HEVC includes some new features which allow high-level parallel computing (at a picture or subpicture level), like Wavefront Parallel Processing (WPP) and tiles, and some new features which allow low-level parallel computing (inside the encoding process), like Local Parallel Method [3] which allows parallel motion estimation.

Tiles are rectangular regions of a video frame which can be independently encoded. In this work we analyze the parallel behavior when tiles are used and how tile partition layout affects both the speed-up and the Rate/Distortion (R/D) performance.

Some works like [4] are focused on parallelizing the decoding side of HEVC, since they are looking for fast decoding of pre-encoded multimedia content, like digital cinema and video on demand. However, we think that more effort should be provided to the encoder side, where the highest computatioonal complexity is found. Parallelizing the video encoding part can be useful for applications like video recording and live event streaming.

There are works that examine and propose low-level parallel techniques for encoding video sequences with HEVC, like the parallelization of the motion estimation module [5] and the parallelization of the intra prediction module [6]. Our work is based in high-level parallel techniques, by using tiles to take advantage of shared memory architectures.

In [7] authors compare slices and tiles encoding performance in HEVC. They show their results in terms of percentage of bit rate increase/decrease. In our study we evaluate tiles performance but we will focus on both complexity reduction related with the encoding process and R/D performance. Even more, we evaluate the impact of the tile partitioning.

The rest of the paper is organized as follows. In Section 2 we will present the main aspects of tiles in HEVC. In Section 3 the results of our tests will be presented and analyzed. At last, several conclusions will be drawn in Section 4.

2 Tiles partitioning

The division of a video frame into tiles is a new technique, included in HEVC, which did not exist in previous video coding standards. Tiles are rectangular regions of a video frame which can be independently encoded (and subsequently decoded), with the use of some common data for the whole frame. This independence allows the parallelization of the encoding and decoding processes at a subpicture level. Tiles contain an integer number of Coding Tree Units (CTUs). Each rectangular region results from the division of a frame into several columns and rows. The width of each column (in CTU units) and also the height of each row can be set individually. In the example shown in Figure 1, a full-HD (1920x1080) frame is divided into 10 tiles using a partitioning scheme of 5 columns with a width of 6 CTUs and 2 rows with a height of 8 and 9 CTUs, respectively.

The independence of tiles allows the parallelization of the encoding process but it has a main drawback: coding efficiency, regarding R/D performance, is reduced. This happens because tiles cannot use information from CTUs belonging to other nearby tiles to make any kind of prediction, and thus, the existing redundancy between nearby CTUs which belong to different tiles cannot be exploited.

In this work we have implemented a tile-level parallelization of the HEVC video encoder for shared memory platforms. The encoding of each tile is assigned to a different core. We

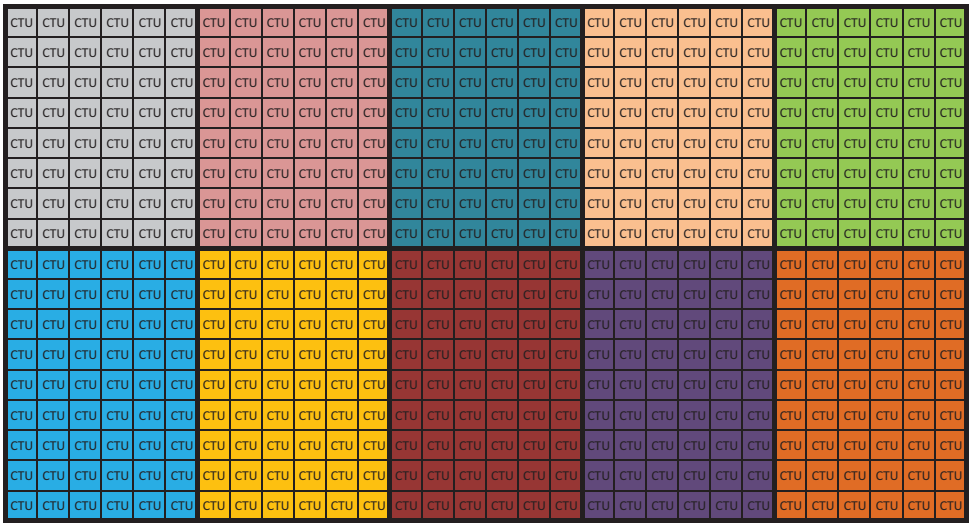


Figure 1: Division of a full-HD frame (1920x1080 pixels) into 10 tiles (5 columns with a width of 6 CTUs; 2 rows with a height of 8 and 9 CTUs each)

have evaluated the performance of the parallel version of the video encoder for 2, 4, 6, 8, 9, and 10 processes and compared the obtained results with those provided by the sequential version, both regarding R/D performance and computing performance. As stated before, we map the tiles per frame onto the same number of processes. For a certain number of tiles per frame, we can find different frame partitions. For example, if we want to divide the frame into 9 tiles we can choose three main different distributions: 9x1 (9 columns by 1 row), 1x9 (1 column by 9 rows), and 3x3 (3 columns by 3 rows). Each one of these distributions can be further designed in different ways if we change the width and height of each one of the columns and rows. In order to get the maximum parallel computing efficiency we will use the layouts that provide the most balanced load distribution, i.e., producing tiles with equal or similar number of CTUs. A balanced load distribution does not always guarantee a balanced work distribution because the resources needed to encode a single CTU may vary, but a completely unbalanced load distribution will likely bring a low parallel computing efficiency. As a measure of the load distribution balance we have calculated the maximum theoretical parallel efficiency, considering the same computational complexity for each CTU. A value of 100% denotes that all the processes will encode the same number of CTUs (and this number is exactly the average value). In Table 1, we have enumerated the different tile partitions that we will use in our tests for different number of cores and two video formats.

The optimum theoretical parallel efficiency would be provided by the "balance %"

Table 1: Layouts and percentage of load balance

(a) 2560x1600 (40x25 CTUs)

#Proc	Layout	AvgCTU	MaxCTU	Bal %
1P	1x1	1000	1000	100%
2P	1x2	500	520	96%
	2x1	500	500	100%
4P	1x4	250	280	89%
	2x2	250	260	96%
	4x1	250	250	100%
6P	1x6	166.7	200	83%
	2x3	166.7	180	93%
	3x2	166.7	182	92%
	6x1	166.7	175	95%
8P	1x8	125	160	78%
	2x4	125	140	89%
	4x2	125	130	96%
	8x1	125	125	100%
9P	1x9	111.1	120	93%
	3x3	111.1	126	88%
	9x1	111.1	125	89%
10P	1x10	100	120	83%
	2x5	100	100	100%
	5x2	100	104	96%
	10x1	100	100	100%

(b) 1920x1080 (30x17 CTUs)

#Proc	Layout	AvgCTU	MaxCTU	Bal %
1P	1x1	510	510	100%
2P	1x2	255	270	94%
	2x1	255	255	100%
4P	1x4	127.5	150	85%
	2x2	127.5	135	94%
	4x1	127.5	136	94%
6P	1x6	85	90	94%
	2x3	85	90	94%
	3x2	85	90	94%
	6x1	85	85	100%
8P	1x8	63.8	90	71%
	2x4	63.8	75	85%
	4x2	63.8	72	89%
	8x1	63.8	68	94%
9P	1x9	56.7	60	94%
	3x3	56.7	60	94%
	9x1	56.7	68	83%
10P	1x10	51	60	85%
	2x5	51	60	85%
	5x2	51	54	94%
	10x1	51	51	100%

column in Table 1. If we divide, for example, a frame with 2560x1600 pixels into 10 tiles, and choose the 10x1 layout, then we will have 10 tiles of 100 CTUs each, which means 100% of load balance (all tiles have the same number of CTUs). If we, instead, select the 1x10 layout, then we will have 5 tiles with 80 CTUs each, and 5 tiles with 120 CTUs each. The most probable scenario is that the 5 processes managing the "small" tiles remain idle waiting for the processes in charge of the "big" tiles. In this case, a maximum load balance index of 83% would be achieved. So, for a specific number of processes, the selected layout may affect the parallel efficiency. Note also that a single layout can provide different load balance percentages depending on the resolution of the video sequence. For example, the 4x1 layout obtains 100% and 94% of load balance for 2560x1600 and 1920x1080 video resolutions, respectively.

In the next section we will present the results of encoding the selected video sequences by using all the layouts presented in Table 1.

3 Numerical experiments

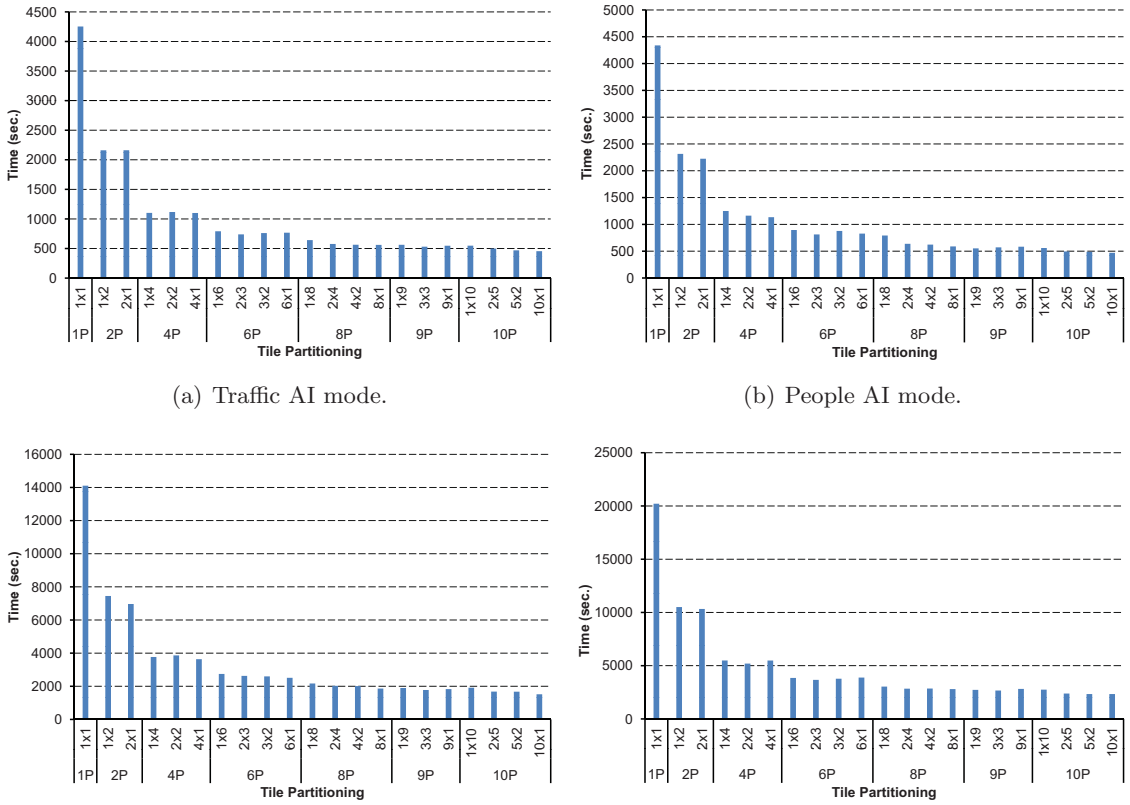
The proposed parallel algorithm has been tested on a shared memory platform consisting of two Intel XEON X5660 hexacores at up to 2.8 GHz and 12MB cache per processor, and 48 GB of RAM. The operating system used is CentOS Linux 5.6 for x86 64 bit. The parallel environment has been managed using OpenMP [8]. The compiler used is *g++* compiler *v*.4.1.2. The reference encoder software used is HM 16.3 [9].

The testing video sequences used in our experiments are Traffic (2560x1600), People on Street (2560x1600), Tennis (1920x1080), and Park Scene (1920x1080), and we present results using Low-delay B (LB) and All Intra (AI) coding modes, encoding 150 frames for Traffic and People sequences and 240 frames for ParkScene and Tennis sequences at different Quantization Parameters (QPs) (22, 27, 32, 37).

In Figure 2, we present the encoding time evolution for Traffic and People sequences in both AI and LB modes as a function of the number of processes with the proposed tile partitioning. As can be seen, the encoding time can be reduced up to 9.3 times when we use 10 processes. As can be seen in Figure 3, the tile-level parallelization algorithm obtains a good parallel performance and and also nice scalability results. Looking at Figure 3, for 10 processes, there are differences in the parallel performance when we use different tile partitioning layouts. Specifically, in AI coding mode we can find differences of up to 1.58x from a tile partitioning scheme of 1x10 respect to the tile partitioning scheme of 10x1. In general, tile partitioning layouts based on columns of CTUs or square tiles obtain better parallel performance. As it would be expected, this effect will depend on the video resolution. Following with the previous example, for a video resolution of 2560x1600, and a CTU size of 64x64, the number of CTUs in a frame are 40x25. If we divide the frame with the 1x10 layout we have 5 processes with 40x2 CTUs and 5 processes with 40x3 CTUs. On the other side, if we divide the frame with the 10x1 layout we have 10 processes with 4x25 CTUs. In the first case (1x10), 5 processes have to perform a 50% more work than the other 5 proccesses. Usually the more balanced the computational load is, the better parallel performance is achieved, except for some sequences where this is not accomplished. In those exceptions, even when each process has the same number of CTUs, the computational complexity inherent to each CTU differs, producing that some processes finish before the others.

In Figure 4 we show the average parallel efficiency for all QP values of the different tested images encoded in both LB and AI modes. As can be seen good efficiencies are obtained for both LB and AI encoding modes, being the efficiency 87% on average. However, if we focus on square tile partitioning layouts, the average parallel efficiency rises till 91%.

Regarding R/D behavior, in Figure 5 we present the % of BD-rate evolution for 4K video sequences as a function of the number of processes and the tile partitioning scheme. As can be seen, the % of BD-rate increases as the number of processes does. This is an expected behavior because tiles are independent structures and therefore, the arithmetic



(a) Traffic AI mode.

(b) People AI mode.

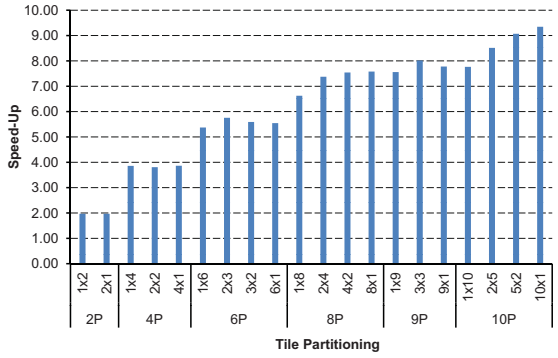
(c) Traffic LB mode.

(d) People LB mode.

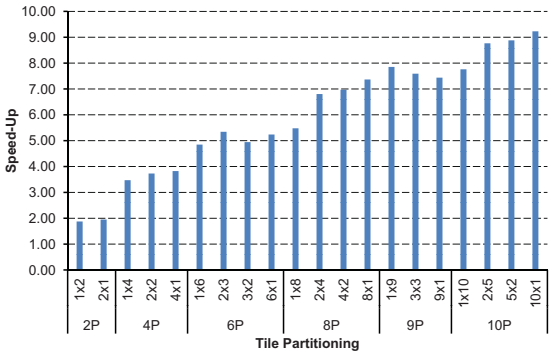
Figure 2: Encoding time evolution for all video sequences with different number of processes and tile partitioning for QP=37

encoder works in an independent way on each tile and no information of previously encoded tiles is available. Furthermore, square tile partitioning performs slightly better in both LB and AI modes, because more information of neighbouring CTUs is available for inter and intra prediction.

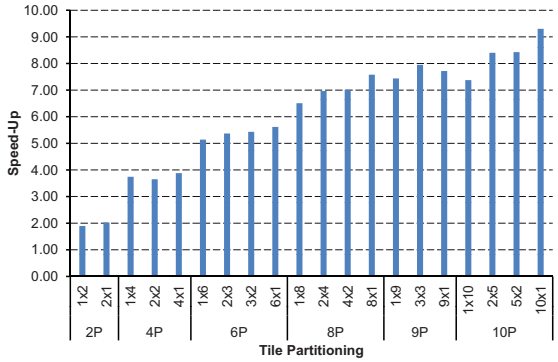
Looking at the results, we can asses that parallelization of HEVC in tiles is worth the effort, because it drastically reduces the encoding time (up to 9.3x for 10 processes) with a low R/D increment, specially if square partitioning schemes are used (0.75% BD-rate for AI mode and 1.2% for LB mode on average). The maximum increment due to tile partitioning scheme is 5% BD-rate increment for Tennis sequence in LB mode using 10 processes. So, the main aspect that will affect the parallel performance is the computational load and this is why we should divide tiles in such a manner that all processes have the same number



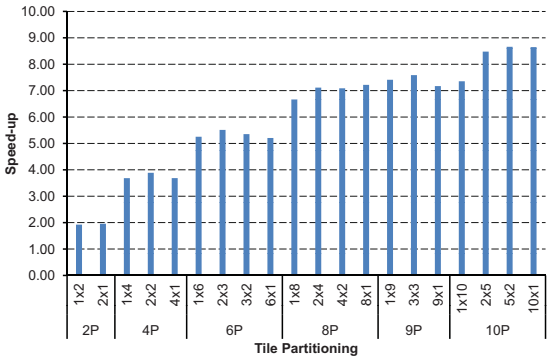
(a) Traffic AI mode.



(b) People AI mode.



(c) Traffic LB mode.



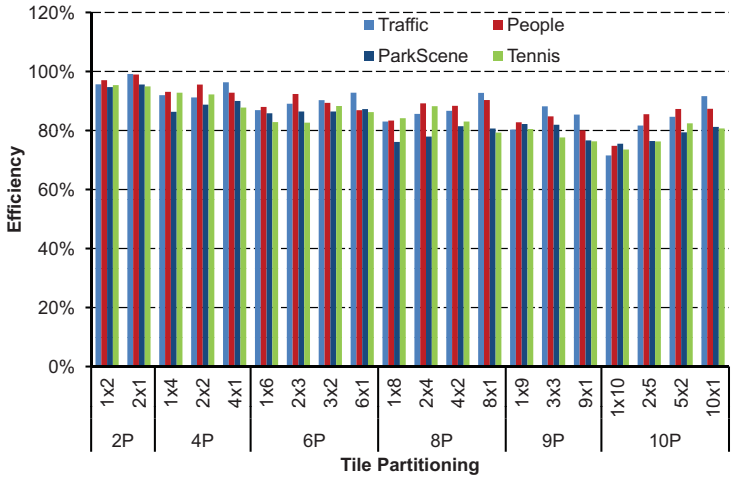
(d) People LB mode.

Figure 3: Speed-Up evolution for all video sequences with different number of processes and tile partitioning for QP=37

of CTUs. In the case of the video sequences analyzed in this work, tiles using square partitioning or tiles with more columns of CTUs than rows, have a better computational load distribution.

4 Conclusions

In this paper we present an study of the tile-level parallelization approach of HEVC encoder when different tile partitioning layouts are used. Results show that both square tile and column tile partitioning layouts obtain the best speed-ups (up to 9.3x for 10 processes) in the tested video sequences. Although in some experiments column-based tile partitioning obtain better parallel efficiency, on average, square tile partitioning layouts present a better



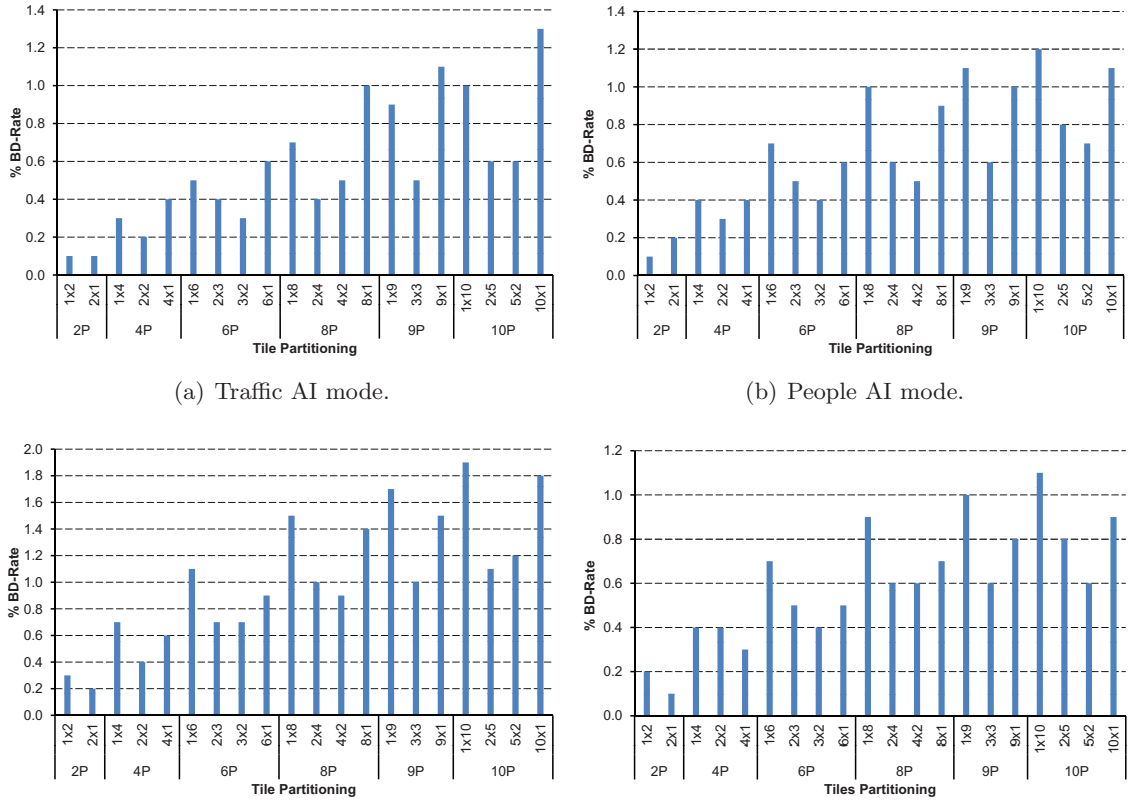
(a) LB mode.

(b) AI mode.

Figure 4: Average parallel efficiency for all tested video sequences with different number of processes and tile partitioning for all QPs

behavior in both speed-up and R/D. Besides, the increment in BD-rate percentage is low in all cases, specially when square tile partitioning is applied, because more information of neighboring CTUs is available for the inter and intra prediction processes.

Finally, we should take into account the video sequence resolution in order to perform



(a) Traffic AI mode.

(b) People AI mode.

(c) Traffic LB mode.

(d) People LB mode.

Figure 5: Average % BD-Rate evolution for all video sequences with different number of processes and tile partitioning for all QPs

the tile partitioning in such a way that the number of CTUs on each tile will be nearly the same and thus the computational load will be more balanced, obtaining better speed-ups.

Acknowledgments

This research was supported by the Spanish Ministry of Economy and Competitiveness under Grant TIN2015-66972-C5-4-R.

References

[1] B. Bross, W.-J. Han, J.-R. Ohm, G. J. Sullivan, Y.-K. Wang, and T. Wiegand, "High Efficiency Video Coding (HEVC) text specification draft 10," Joint Collaborative Team on Video Coding (JCT-VC), Geneva (Switzerland), Tech. Rep. JCTVC-L1003, January 2013.

[2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," *ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC) version 16, 2012*, 2012.

[3] M. Zhou, "AHG10: Configurable and CU-group level parallel merge/skip," Joint Collaborative Team on Video Coding-H0082, Tech. Rep., 2012.

[4] M. Alvarez-Mesa, C. Chi, B. Juurlink, V. George, and T. Schierl, "Parallel video decoding in the emerging HEVC standard," in *International Conference on Acoustics, Speech, and Signal Processing, Kyoto*, March 2012, pp. 1–17.

[5] Q. Yu, L. Zhao, and S. Ma, "Parallel AMVP candidate list construction for HEVC," in *VCIP'12*, 2012, pp. 1–6.

[6] J. Jiang, B. Guo, W. Mo, and K. Fan, "Block-based parallel intra prediction scheme for HEVC," *Journal of Multimedia*, vol. 7, no. 4, pp. 289 –294, August 2012.

[7] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, and M. Zhou, "An overview of tiles in HEVC," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 6, pp. 969–977, Dec 2013.

[8] "Openmp application program interface, version 3.1," *OpenMP Architecture Review Board. http://www.openmp.org*, 2011.

[9] HEVC Reference Software, https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.3/.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Tumour–immune system interaction model with Erlang distributed delays

Monika Joanna Piotrowska[1] **and Marek Bodnar**[1]

[1] *Faculty of Mathematics, Informatics and Mechanics, Institute of Applied Mathematics and Mechanics, University of Warsaw*

emails: monika@mimuw.edu.pl, mbodnar@mimuw.edu.pl

Abstract

In the presented paper we consider generalised tumour-immune system interaction model with distributed delays. Existence, uniqueness and non-negativity of solutions of considered system are proved. Next, the existence of the steady states is discussed. For the particular choice of Erlang probability densities functions stability of steady states is analytically investigated. Moreover, the model is fitted to the experimental data. For the fitted values of the parameters the influence of model parameters on the value of the critical average delay that destabilises the steady state is studied numerically. Next, the influence of the two most important parameters on the model dynamics is discussed.

Key words: tumour-immune system interactions, distributed delay, Hopf bifurcation

1 Introduction

In [11] the model of tumour–immune system interactions that takes into account two most important variables: the size of specific anti-tumour immunity (X) and the size of tumour or to be precise amount of tumour antigens (Y) is considered. The dynamics is governed by two ordinary differential equations with discrete delay:

$$\frac{\mathrm{d}X(s)}{\mathrm{d}s} = w - uX(s) + a_F F(Y(s-\tau))X(s) - bX(s)Y(s), \qquad \frac{\mathrm{d}Y(s)}{\mathrm{d}s} = Y(s)(r - cX(s)). \tag{1}$$

In the model it is assumed that in the absence of tumour antigens there is a constant production (with a rate w) of precursor cells that are able to respond to the tumour antigens while the natural immune cell death is modelled by a linear term uX. Thus, in the absence of tumour cells a constant level

 ISBN: 978-84-608-6082-2

of immunity (called background immunity or native immunity) w/u is kept within the body. The presence of tumour antigens initiate an increase of immunity's size, which growth rate is assumed to be proportional to the current size of the immunity X and to the coefficient of proportionality $(a_F F)$, which most often, [2, 10, 9, 6], is assumed to be only a bounded function of tumour antigens i.e. $F = F(Y) = Y^\alpha/(k^\alpha + Y^\alpha)$, often with $\alpha = 1$. The immune reaction on the detection of antigen requires transmitting signals that initiates production of appropriate immune agents, production of certain cells and proteins, etc. thus this process takes some time (denoted by τ). To reflect that phenomena and following [5] a discrete time lag was incorporated into the system in the stimulus function.

Tumour cells may produce factors that decrease the activity of immune system e.g. by a chemical process of antigen-antibody interaction [6]. This process is modelled by a bi-linear term XY with a constant coefficient b. Moreover, it is assumed that change of tumour size depends on two processes: cells growth, which is exponential with a constant rate r [13, 1, 12], and destructive influence of immune system on tumour cells modelled, similarly as in [4], by a bi-linear term with a constant rate c. Due to the biological interpretation of constants, it is assumed that all parameters are non-negative, and in particular parameters a, c, u, r, w and k are strictly positive. Moreover, in[6] it is assumed that for parameter α, describing the switching characteristics of the stimulus functions, $\alpha \geq 1$ holds.

Of course, in reality the duration of a process is never constant and it usually fluctuates around some value. Thus, it is believed that delay is distributed around some average values instead of being concentrated in points. Therefore, here we study system (1), in which instead of discrete delays we consider delay distributed around some mean values with distributions given by general probability densities $K(s) : [0, \infty) \to \mathbb{R}_0^+$. Hence, the considered system, in the non-dimensional form reads

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = d\Big(\theta - x(t) + \rho x(t) \int_0^\infty K(s) f(y(t-s))\,\mathrm{d}s - \psi x(t) y(t)\Big), \qquad \frac{\mathrm{d}y(t)}{\mathrm{d}t} = y(t)(1 - x(t)), \tag{2}$$

where $0 \leq \tau_{\mathrm{av}} = \int_0^\infty sK(s)ds < \infty$ and $\int_0^\infty K(s)ds = 1$ and f fulfils

(A1) f is a non-negative $\mathbf{C}^2$ class function defined on $\mathbb{R}_{\geq} = [0, +\infty)$,

(A2) $f(0) = 0$, $\lim_{y \to +\infty} f(y) = 1$, $f(1) = 1/2$; f is increasing on $\mathbb{R}_{\geq}$;

(A3) f has at most one inflection point.

In fact, so high smoothness assumed in (A1) is required only for the bifurcation results. For the global existence and uniqueness results this can be weaken assuming only a Lipschitz continuity of the function f. Nevertheless, in our opinion from the point of view of applications Assumption (A1) is not very restrictive. Assumptions (A2)–(A3) ensure that the function f has the Michaelis-Menten or "s" shape as a generalisation of the functions considered in [2, 10, 6, 11]. Assumption (A3) is essential for the results regarding the existence of positive steady states. If we relax this condition, the system considered in this paper can have arbitrary number of steady states. On the other hand, if

f has no inflection point, there can exist at most two positive steady states, while for f having one inflection point, there can exist up to three positive steady states.

Due to technical reasons, we choose the following Banach space

$$\mathcal{K} = \left\{ \varphi \in C : \lim_{\xi \to -\infty} \varphi(\xi)\, \mathrm{e}^{\xi} = 0, \sup_{\xi \in (-\infty, 0]} |\varphi(\xi)\, \mathrm{e}^{\xi}| < \infty \right\}, \ \|\varphi\| = \sup_{\xi \in (-\infty, 0]} |\varphi(\xi)\, \mathrm{e}^{\xi}| \text{ for } \varphi \in \mathcal{K},$$

where C denotes the space of continuous functions defined on $(-\infty, 0]$. Let $\mathcal{K}_{\geq}$ denotes the subspace of non-negative functions belonging to $\mathcal{K}$. Thus, we consider initial conditions from $\mathcal{K}_{\geq}$.

2 Model analysis

The right-hand side of system (2) fulfils local Lipschitz condition. Thus, the there exists a unique solution to system (2) defined on $t \in [0, t_{\varphi})$ for some $t_{\varphi} > 0$ (see [7, Chapter 2, Theorem 1.2]). The non-negativity of solutions is a consequence of the form of the right hand side of system (2). Moreover, using rather standard estimations we are able to show that the solution of (2) can be prolonged for $t \geq 0$. Hence the following holds

Theorem 2.1 (global existence and uniqueness) *Let f fulfil conditions* (A1)–(A2)*. Then for any initial data $(\varphi_1, \varphi_2) \in K_{\geq}$ there exists a unique solution of* (2) *in $\mathcal{K}_{\geq}$ defined on $t \in [0, +\infty)$.*

System (2) has one semi-trivial steady state and from zero up to three positive steady states as summarised in Table 1. Notice that all positive steady stats have form $(1, \zeta)$, where ζ is a positive constant.

Theorem 2.2 *Let f fulfils conditions* (A1)–(A2)*, then the semi-trivial steady state $A = (\theta, 0)$ of system* (2) *is a locally stable node for $\theta > 1$ and a saddle for $\theta < 1$ independently of the form of the delay probability density K.*

Proof : Linearising system (2) around the steady state A and due to the assumption (A2) we obtain

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -dx(t) - d\psi\theta y(t) + d\rho\theta f'(0) \int_0^{\infty} K(s) y(t-s)\, \mathrm{d}s, \qquad \frac{\mathrm{d}y(t)}{\mathrm{d}t} = (1-\theta) y(t).$$

The characteristic function reads $W(\lambda) = (\lambda + d)(\lambda - (1-\theta))$. Since it does not depend on the delay, the stability of the steady state A does not depend on delay. ■

To study the local behaviour of solutions around the steady state $(1, \zeta)$ we linearise system (2) obtaining

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -dx\theta + d\gamma_f \int_0^{\infty} K(s) y(t-s)\, \mathrm{d}s - d\psi y(t), \qquad \frac{\mathrm{d}y(t)}{\mathrm{d}t} = -\zeta x(t).$$

 ISBN: 978-84-608-6082-2

Table 1: Existence and stability of the steady states of scaled version of system (1) with discrete delay and f fulfilling conditions (A1)–(A3) depending on θ, ψ, ρ values. In the column ips we indicate the number of possible inflection points of the function f. The Hopf bifurcation occurs whenever a steady state loses its stability due to increase of discrete time delay.

ψ	θ	Possible steady states	ips	Stability
		Semi-trivial steady states		
$\psi \geq 0$	$0 < \theta < 1$	$A = (\theta, 0)$	0,1	unstable for all $\tau \geq 0$
	$\theta > 1$	$A = (\theta, 0)$		locally stable for all $\tau \geq 0$
		Non-trivial positive steady states		
$\psi = 0$	$\max\{0, 1-\rho\} < \theta < 1$	$R = (1, y_R)$, $y_R > 0$	0,1	globally stable for $\tau = 0$ locally stable for $\tau < \tau^{cr}_{y_R}$
	$\theta > 1$ or $0 < \theta < 1 - \rho$	none		
$\psi > 0$	$0 < \theta < 1$	none	0,1	
		$B = (1, y_B)$, $C = (1, y_C)$ $0 < y_B < y_C$	0,1	B locally stable for $\tau < \tau^{cr}_{y_B}$ C unstable for all $\tau \geq 0$
	$\theta > 1$	$C = (1, y_C)$, $y_C > 0$	0,1	C unstable for all $\tau \geq 0$
		$B = (1, y_B)$, $C = (1, y_C)$, $D = (1, y_D)$ $0 < y_B < y_C < y_D$	1	B unstable for all $\tau \geq 0$ C locally stable $\tau < \tau^{cr}_{y_C}$ D unstable for all $\tau \geq 0$

Then the characteristic function has the following form

$$W(\lambda) = \lambda^2 + d\theta\lambda - d\psi\zeta + d\zeta\gamma_f \int_0^\infty K(s)\, \mathrm{e}^{-\lambda s}\, \mathrm{d}s, \qquad \gamma_f(\zeta) = \rho f'(\zeta), \tag{3}$$

where for γ_f the dependence on ζ is omitted for simplicity of notation.

For the positive steady states stability analysis we consider Erlang probability densities

$$K(s) = \frac{a^m (s)^{m-1}}{(m-1)!}\, \mathrm{e}^{-a(s)},$$

where $a \in \mathbb{R}^+$, $m_i \in \mathbb{Z}^+$, $s \in \mathbb{R}_0^+$. Thus, $\int_0^\infty K(s)\, \mathrm{e}^{-\lambda s}\, \mathrm{d}s = a^m/(a+\lambda)^m$ and the characteristic function reads

$$W(\lambda) = \lambda^2 + d\theta\lambda - d\psi\zeta + d\zeta\gamma_f \frac{a^m}{(a+\lambda)^m}.$$

Hence, in the following we consider

$$D(\lambda) = (a+\lambda)^m(\lambda^2 + d\theta\lambda - d\psi\zeta) + d\zeta\gamma_f a^m. \tag{4}$$

Theorem 2.3 *Let f fulfils conditions* (A1)–(A3)*. Consider system* (2) *with probability density given by* (2)*, then*

 ISBN: 978-84-608-6082-2

(i) *for $m = 1$ the steady state $(1, \zeta)$ is stable if $\gamma_f > \psi$ and $a > a_{cr}$, and it is unstable if $\gamma_f < \psi$ or $a < a_{cr}$. For $\gamma_f > \psi$, the Hopf bifurcation occurs at*

$$a_{cr} = \frac{\zeta\gamma_f - \theta^2 d + \sqrt{(\theta^2 d - \gamma_f\zeta)^2 + 4\theta^2\psi\zeta d}}{2\theta}. \tag{5}$$

(ii) *for $m = 2$ the steady state $(1, \zeta)$ is stable if $a > a_{cr}$ and $\gamma_f > \psi$, where a_{cr} is a unique solution of $2\theta(d\psi\zeta - a(a + d\theta))^2 = a(2a + d\theta)^2\gamma_f$, with $a \geq 2\psi\zeta/\theta$. For $a < a_{cr}$ or $\gamma_f < \psi$, the steady state $(1, \zeta)$ is unstable, while for $\gamma_f > \psi$, the Hopf bifurcation occurs at $a = a_{cr}$.*

The proof of Theorem 2.3 is based on the study of the existence of roots of a polynomial (which form changes depending on the considered case) using the Descartes' rule of signs, mathematical analysis and Theorem 1 from [3].

Clearly, for given parameters the threshold value a_{cr} can be calculated from the Routh-Hurwitz stability criterion. However, already for $m = 2$, the formula for the threshold value a_{cr} is implicit. For $m > 2$, the characteristic polynomial is of degree at least five and the Routh-Hurwitz condition consist of at least two inequalities that can not be, in general, reduced to a simpler form. Thus, we decided not to write these inequalities for $m > 2$.

On the other hand we know that for $\gamma_f - \psi < 0$ the steady state $(1, \zeta)$ of the model without delay is unstable [6, 11]. Thus, it remains unstable also for system (2) with Erlang probability density (2). Moreover, using the continuity argument we are able to show the following

Theorem 2.4 *Let f fulfils conditions (A1)–(A3). If $\gamma_f > \psi$, then for Erlang distribution probability density there exist $0 < \bar{a}_1 \leq \bar{a}_2$ such that the steady state $(1, \zeta)$ is unstable for $a \in (0, \bar{a}_1)$ and it is stable for $a > \bar{a}_2$.*

3 Numerical simulations

In our numerical study we limit ourselves to a specific form of the function f, namely $f(y) = y/(1 + y)$ and Erlang distributions. In [11] model parameters were estimated by the fitting procedure of model (1) trajectories to two sets of the experimental data (B-cell lymphoma in the spleen tumour of normal and chimeric mice) from [14]. Fitting procedure was performed in Matlab using particle swarm optimization algorithm and standard mean square error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} Err_i^2, \qquad Err_i = (x_i^{data} - x_i^{sym})/x_i^{data},$$

where n is the number of experimental measurements, x_i^{data} are the measured values and x_i^{sym} are the values of the x solution of the considered system for respective time points.

 ISBN: 978-84-608-6082-2

The initial conditions reflected the modelled experiments that is injection of 5×10^5 cells at time 0 and the initial immune system equal to the background immunity for $t \in [-\tau, 0]$. In [11] first parameters were fitted to the chimeric mice experiment including the initial condition, with MSE at the level of 3.57%. Then part of parameters determining the initial condition (that is w, u) and parameter r (tumour growth rate) were fixed and the rest of the parameters were fitted to the control experiment with MSE at the level of 0.14%, for scaled values of model parameters, see Table 2. Assuming the same as in [11] ranges of the parameters we follow the procedure described above however this time using linear chain trick [8] with appropriate initial conditions.

Table 2: Values of the fitted (scaled) parameters for system (2), $f = y/(1+y)$.

m		θ	d	ρ	ψ	a	$\tau_{av}\backslash\tau$	MSE
1	control	0.5744	0.0412	7.4029	1.5277	0.6722	1.4876	0.21%
	chimeric	0.6872		13.6316	5.8918	38.6986	0.0258	3.57%
2	control	0.5778	0.0415	7.4567	1.4974	1.2870	1.5540	0.13%
	chimeric	0.6873		13.7073	5.9061	66.5506	0.0301	3.58%
5	control	0.5783	0.0419	7.3589	1.4596	3.3367	1.4985	0.12%
	chimeric	0.6881		14.2413	6.1840	110.2895	0.0453	3.61%
∞	control	0.5802	0.0418	7.4046	1.4576	—	1.4964	0,14%
	chimeric	0.6841		14.4583	6.3802	—	0.0273	3.57%

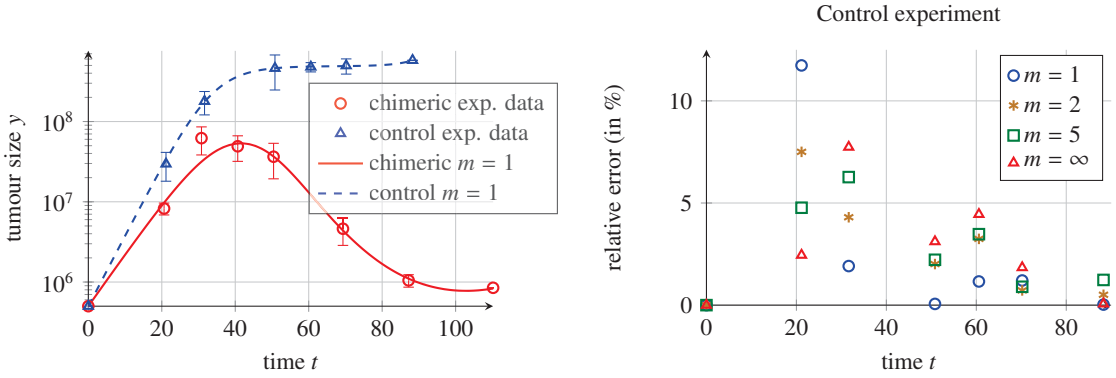


Figure 1: (Left) Result of the fitting procedure for Erland distribution with $m = 1$ for system (2). (Right) Comparison of the relative errors ($|\mathrm{Err}_i|$) for the control experiment for different Erlang distributions as described in the legend, $m = \infty$ represents data for the model with discrete delay.

Fitting results for the particular Erlang distributions, showing similar fitting accuracy, are presented in Table 2. For better understanding, in Fig. 1 we present comparison of the relative errors for each measurement point ($|\mathrm{Err}_i|$) for the control and chimera experiments for different Erlang dis-

tributions. In addition, in Table 3 we give the coordinates of non-negative steady states, determine theirs stability and critical average delay for which steady states B loose the stability.

Next, for the fitted sets of parameters we investigate the dependence of critical average delay for the positive steady states (which my loses stability) on the model parameters θ and ρ, see Figs. 2 and 3, showing that plotted curves are located very close to each other. The biological/medical justification of the parameter choice is described in the next section.

Table 3: Steady states and theirs stability for system (2) for the parameters as given in Tab. 2.

m		$(x_A, 0)$ unstable	$(1, y_B)$ unst./st.	$(1, y_c)$ unstable	a_{cr}	$\tau_{av}^{cr} \backslash \tau_{y_B}^{cr}$
1	control	(0.5744, 0)	(1, 0.0799)	(1, 3.4874)	0.8650	1.1560
	chimeric	(0.6872, 0)	(1, 0.0436)	(1, 1.2170)	0.7798	1.2823
2	control	(0.5778, 0)	(1, 0.0779)	(1, 3.6199)	1.7121	1.1682
	chimeric	(0.6873, 0)	(1, 0.0432)	(1, 1.2247)	1.5695	1.2743
5	control	(0.5783, 0)	(1, 0.0786)	(1, 3.6743)	4.2743	1.1698
	chimeric	(0.6881, 0)	(1, 0.0417)	(1, 1.2108)	3.9516	1.2653
∞	control	(0.5802,0)	(1, 0.0775)	(1, 3.7145)	—	1.1776
	chimeric	(0.6841,0)	(1, 0.0422)	(1, 1.1745)	—	1.2227

4 Results and Discussion

For each considered Erlang distribution we were able to fit the solutions of the system to the experimental data with the MSEs at the same comparable level. For fitted sets of parameters, in all cases, system (2) has three steady states one semi-trivial (unstable) and two positive from which one is unstable (for both chimera and control experiments) and the second one (dormant steady state), is stable for chimera experiment and unstable for control experiment.

We investigated the influence of the parameters θ and ρ on the stability of the steady state. The parameter θ was chosen because it is proportional to the constant influx of effector cells w and the killing effectiveness of effector cells c. The ratio c that can be up-regulated e.g. by the infusions of the cells grown ex vivo or as the result of the infusion of properly chosen cytokines together with active in vivo immunisation against the particular tumour cell types. The parameters θ is also inversely proportional to the natural lymphocyte death rate (u) and the tumour growth rate (r) but they are rather out of reach of immunotherapy. Thus, we investigated the dependence of the critical average delay τ_{av}^{cr} (for steady state B) on θ parameter showing that the control experiment the largest stability region we obtain for Erlang distribution with $m = 1$ (smaller for $m = 2$ and $m = 5$) and the smallest for discrete delay. For the chimera experiment situation is is similar, see Fig. 2.

On the other hand, the parameter ρ reflects the stimulation strength and it is proportional to the multiplication of the immunity after the stimulation by antigen and reverse proportional to the

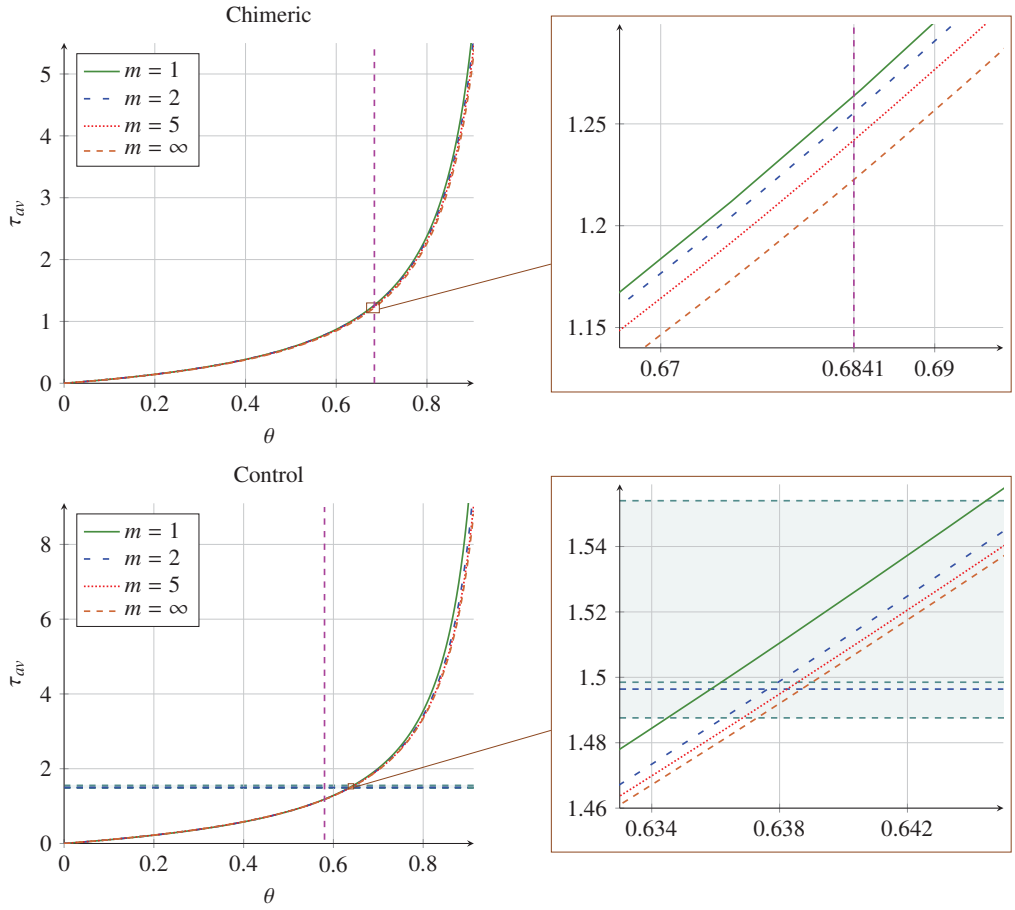


Figure 2: Dependence of the critical average delay τ_{av}^{cr} (for steady state B) on the parameter θ for different values on m in the case of Erlang distributions (as reported in Table 2) and discrete delay for control (top panels) and chimer (bottom panels) sets of parameters. Case $m = \infty$ denotes the results for system (1) with discrete delay. Purple vertical lines indicate the values of θ parameter estimated due to the fitting procedure for discrete case. Blue horizontal dashed lines indicate the values of τ parameter estimated due to the fitting procedure for all considered cases. The stability regions of the steady state B lie below the curves for different values of m.

natural immune cell death rate, that is $\rho = a_F/u$. This parameter has also essential influence on the control of the tumour growth since, together with other parameters, it determines the number of existing steady states and theirs stability. If the tumour antigens are not presented to the lymphocytes, one can increase a_F and possibly speed up the immune response time triggered by the antigens

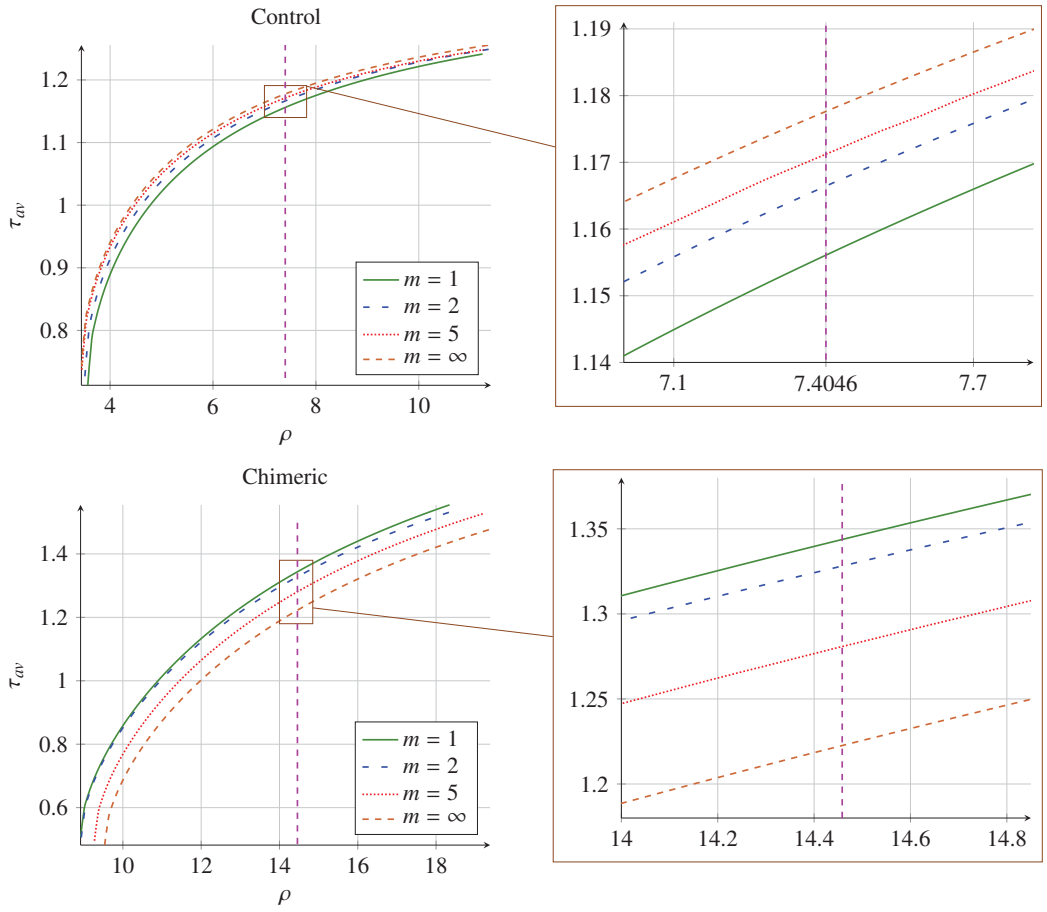


Figure 3: Dependence of the critical average delay τ_{av}^{cr} (for steady state B) on the parameter ρ for different values on m in the case of Erlang distributions (as reported in Table 2) and discrete delay for control (top panels) and chimer (bottom panels) sets of parameters. Case $m = \infty$ denotes the results for system (1) with discrete delay. Purple vertical dashed line indicates the value of the ρ parameter estimated due to the fitting procedure for system (1) with discrete delay. The stability regions of the steady state B lie below the curves for different values of m.

presence by the method of an artificial loading of the antigen-presenting cells with specific tumour antigens. Changing ρ for control experiment we observe that the largest stability region is obtained for discrete delay and smallest for Erlang distribution with $m = 1$. For the chimeric experiment situation strongly differs – largest stability region is obtained for Erlang distribution with $m = 1$ and smallest for the system with discrete delay.

Clearly, for all considered cases we see that an increase of both θ and ρ increases the value of the average critical delay for which the Hopf bifurcation is observed, thus it enlarges the stability region for the positive dormant steady state B. However, comparing the plots in Fig. 2 and values given in Tables 2 and 3 we conclude that changing the parameter θ in the biologically reachable range and keeping other parameters fixed we are unable to stabilize the dormant steady state for the control experiment. On the other hand, performed simulations partially presented in Fig. 3) show that stabilisation of the dormant steady state for the control is impossible when the parameter ρ (keeping other parameters fixed) is changed. It is so due to the fact that for $\rho = 200$ (maximal biologically justified value of that parameter) the critical value of average delay are: 1.3941 1.3912, 1.3839, 1.3897 for Erlang distributions with $m = 1$, $m = 2$, $m = 5$ and for the discrete delay case, respectively. All these values are smaller than the fitted average delays. Moreover, numerical simulations suggest that this critical value does not exceeds 1.4 even for very large ρ.

Clearly, due to the Linear Chain Trick, the sequence of linear reactions with the same coefficients described by sufficiently large system of ODEs can be modelled by the system with Erlang distribution in which the effect of reaction sequence is expressed by a proper integral. However, assuming more complex sequences of reactions (e.g. containing parallel feedback, coherent feed-forward loops or combination of both) and proceeding in a similar manner as a result one would not obtain such easily treated probability densities.

On the other hand, the solutions of system with discussed probability densities fit well to all experimental measurements except one point for the chimera experiment. This could be due to model limitation, measurement mistake or it might suggest to consider other delay distribution, e.g. piecewise linear. However, that would require the usage of different mathematical tools due to the form of the characteristic function. We will address that these issues in the future work.

Acknowledgements

This work has been partially supported by Ministry of Science and Higher Education Republic of Poland within the Iuventus Plus Grant: “Mathematical modelling of neoplastic processes” (grant No. IP2011041971) from the budget for science in years the 2012-2014.

References

[1] S. BANERJEE AND R. SARKAR, *Delay-induced model for tumor-immune interaction and control of malignant tumor growth*, BioSystems, 91 (2007), pp. 268–288.

[2] G. BELL, *Predator-pray equation simulating an immune response*, Mathematical Biosciences, 16 (1973), pp. 291–314.

[3] K. L. COOKE AND P. VAN DEN DRIESSCHE, *On zeroes of some transcendental equations*, Funkcj. Ekvacioj, 29 (1986), pp. 77–90.

[4] A. d'Onofrio, *A general framework for modeling tumor-immune system competition and immunotherapy: Mathematical analysis and biomedical inferences*, Physica D: Nonlinear Phenomena, 208 (2005), pp. 220–235.

[5] A. d'Onofrio, F. Gatii, P. Cerrai, and L. Freschi, *Delay-induced oscillatory dynamics of tumour immune system interaction*, Mathematical and Computer Modeling, 51 (2010), pp. 527–591.

[6] U. Foryś, J. Waniewski, and P. Zhivkov, *Anti-tumor immunity and tumor anti-immunuty in a mathematical model of tumor immunotherapy*, J. Biol. Sys., 14(1) (2006), pp. 13–30.

[7] Y. Hino, S. Murakami, and T. Naito, *Functional Differential Equations with Infinite Delay*, vol. 1473 of Lecture Notes in Mathematics, Springer-Verlag, New York, 1991.

[8] Y. Kuang, *Delay differential equations with applications in population dynamics*, Academic Press Inc., 1993.

[9] V. A. Kuznetsov, I. A. Makalkin, M. A. Taylor, and A. S. Perelson, *Nonlinear dynamics of immunogenic tumors: Parameter estimation and global bifurcation analysis*, Bull. Math. Biol., 56 (1994), pp. 295–321.

[10] H. Mayer, K. Zaenker, and U. an der Heiden, *A basic mathematical model of the immune response*, Chaos, 5 (1995), pp. 155–161.

[11] M. J. Piotrowska, *An immune system—tumour interactions model with discrete time delay: Model analysis and validation*, Communications in Nonlinear Science and Numerical Simulation, 34 (2016), pp. 185–198.

[12] D. Rodriguez-Perez, O. Sotolongo-Grau, R. Espinosa Riquelme, O. Sotolongo-Costa, J. A. Santos Miranda, and J. C. Antoranz, *Assessment of cancer immunotherapy outcome in terms of the immune response time features*, Mathematical Medicine and Biology, 24 (2007), pp. 287–300.

[13] O. Sotolongo-Costa, L. Morales Molina, D. Rodriguez Perez, J. Antoranz, and M. Chacon Reyes, *Behavior of tumors under nonstationary therapy*, Physica D: Nonlinear Phenomena, 178 (2003), pp. 242–253.

[14] J. W. Uhr, T. Tucker, R. D. May, H. Siu, and E. S. Vitetta, *Cancer dormancy: Studies of the murine BCL_1 lymphoma.*, Cancer Res. (Suppl.), 51 (1991), pp. 5045s–5053s.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Explicit and efficient exponential splitting time integrator for the Klein-Gordon equation with Absorbing Boundary Conditions

A. M. Portillo[1] and I. Alonso-Mallo[1]

[1] *Departamento de Matemática Aplicada, Universidad de Valladolid, Spain*

emails: anapor@mat.uva.es, isaias@mac.uva.es

Abstract

Klein-Gordon equations on an unbounded domain are considered in one dimensional and two dimensional cases. Numerical computation is reduced to a finite domain by using the Hagstrom-Warburton high-order absorbing boundary conditions. The space discretization is reached by means of finite differences on a uniform grid, with fourth order inside the computational domain. Time integration is made by means of exponential splitting schemes that are efficient and easy to implement. Numerical experiments displaying the accuracy of the numerical solution for the two dimensional case are provided.

Key words: Splitting methods, Absorbing boundary conditions, Auxiliary variables, Artificial boundary, Finite differences.

1 Introduction

We consider dispersive waves propagating in $(-\infty, \infty) \times [a, b]$, a two dimensional strip. The south and north boundaries of the strip are denoted by Γ_S and Γ_N. Inside the strip, we consider the Klein-Gordon equation,

$$\partial_t^2 u - c^2 \nabla^2 u + s^2 u = f. \tag{1}$$

Here $c = c(x, y)$ is the given wave speed, $s = s(x, y)$ the medium dispersion coefficient and $f(x, y, t)$ is a given source.

 ISBN: 978-84-608-6082-2

Our results hold when, outside a compact region Ω_0, the speed c and the dispersion coefficient are constant and the source f vanishes. However, for the sake of simplicity, we only consider examples which satisfy this assumption in the whole strip $(-\infty, \infty) \times [a, b]$.

On the south and north boundaries we consider Neumann boundary conditions,

$$\partial_y u = 0, \quad \text{on}\, \Gamma_S \,\text{and}\, \Gamma_N. \tag{2}$$

Finally, we consider the initial conditions,

$$u(x, y, 0) = u_0(x, y), \quad \partial_t u(x, y, 0) = v_0(x, y), \tag{3}$$

which satisfy the boundary conditions on Γ_S and Γ_N, and vanish outside Ω_0.

The numerical approximations of these problems need to reduce the computation to a finite domain. Therefore, we truncate the infinite domain by introducing the west artificial boundary Γ_W, located at $x = x_W$, $a \le y \le b$, and the east artificial boundary Γ_E at $x = x_E$, $a \le y \le b$. We denote by Ω the computational domain bounded by $\Gamma_N \cup \Gamma_W \cup \Gamma_S \cup \Gamma_E$, such that $\Omega_0 \subset \Omega$.

The function u satisfies the Klein-Gordon equation (1) inside Ω, the Neumann boundary condition (2) on Γ_S and Γ_N, and the initial conditions (3) in Ω. It is necessary to define suitable artificial boundary conditions on the artificial boundaries Γ_W and Γ_E. For this, we have focused on the so called Absorbing Boundary Conditions (ABCs), which are designed to produce small reflections inside the computational domain and to have local character. We have considered the Hagstrom-Warburton high-order ABCs wich use auxiliary variables to avoid high derivatives in their formulation. Arbitrary order of absorption P can be achieved by introducing P auxiliary variables ϕ_j, $j = 1, \ldots, P$, satisfying the recursive relations

$$\begin{aligned}
(\partial_t + c\partial_x)u &= \partial_t \phi_1, \\
\partial_t^2 \phi_1 &= c^2(\frac{1}{2}\partial_y^2\phi_0 + \frac{1}{4}\partial_y^2\phi_1 + \frac{1}{4}\partial_y^2\phi_2) - s^2(\frac{1}{2}\phi_0 + \frac{1}{4}\phi_1 + \frac{1}{4}\phi_2), \\
\partial_t^2 \phi_j &= c^2(\frac{1}{4}\partial_y^2\phi_{j-1} + \frac{1}{2}\partial_y^2\phi_j + \frac{1}{4}\partial_y^2\phi_{j+1}) - s^2(\frac{1}{4}\phi_{j-1} + \frac{1}{2}\phi_j + \frac{1}{4}\phi_{j+1}), j = 2, \ldots, P, \\
u &= \phi_0, \quad \phi_{P+1} = 0.
\end{aligned}$$

2 Spatial discretization

For the sake of simplicity, we consider the same size step in both directions x and y, that is, for a value of N, $h = \dfrac{x_E - x_W}{N}$ and $M = \dfrac{b-a}{h}$. Let $x_j = x_W + (j-1)h$, $j = 1, \ldots, N+1$, and $y_l = a + (l-1)h$, $l = 1, \ldots, M+1$, be the nodes of the spatial discretization. This produces a uniform grid in the computational domain with $M+1$ rows and $N+1$ columns. We denote $u_{jl}(t) = u(x_j, y_l, t)$. In this way, there is a matrix of unknowns. On the other

hand, we consider $\phi_{rl} = \phi_r(y_l)$, $r = 1, \ldots, P$, and $l = 1, \ldots, M+1$, on west and east boundaries.

Second order spatial derivatives in the direction x, $\partial_x^2 u_{jl}$, from $j = 3$ to $N-1$, are approximated by fourth order central finite differences and, for $j = 2$ and N, by fourth order one-sided finite differences. First order spatial derivatives $\partial_x u_{1l}$ and $\partial_x u_{N+1,l}$ are approximated by fourth order one-sided finite differences.

We assume that the unknowns associated with nodes on south and north boundaries have been removed using Neumann boundary conditions. Spatial derivative in the direction y, $\partial_y^2 u_{jl}$, from $l = 3$ to $M-1$, are approximated by fourth order central finite differences. For $l = 2$ and M, fourth order one-sided finite differences and Neumann boundary condition are used to obtain the approximation to $\partial_y^2 u_{j2}$ and $\partial_y^2 u_{jM}$.

3 Time discretization: exponential splitting method

For the time discretization, we propose a fourth order exponential splitting method which improves the computational efficiency of the time integration.

The main idea of splitting methods for the time integration of ordinary differential equations involves to separate the system into several parts, being each of them easily integrable, and then combining the solutions of the intermediate problems to achieve a good approximation of the original problem.

Let it be $\mathbf{u}_j$ the column j, $\boldsymbol{\phi}_j^W$ and $\boldsymbol{\phi}_j^E$ the column vectors corresponding to the auxiliary variable j on west and east boundary respectively. We consider $\mathbf{u} = [\mathbf{u}_1^T, \ldots, \mathbf{u}_{N+1}^T]^T$, $\mathbf{u}' = [\frac{d}{dt}\mathbf{u}_2^T, \ldots, \frac{d}{dt}\mathbf{u}_N^T]^T$, $\boldsymbol{\phi}^W = [(\boldsymbol{\phi}_1^W)^T, \ldots, (\boldsymbol{\phi}_P^W)^T]^T$, $\boldsymbol{\phi}^E = [(\boldsymbol{\phi}_1^E)^T, \ldots, (\boldsymbol{\phi}_P^E)^T]^T$, $(\boldsymbol{\phi}^W)' = [\frac{d}{dt}(\boldsymbol{\phi}_1^W)^T, \ldots, \frac{d}{dt}(\boldsymbol{\phi}_P^W)^T]^T$ and finally $(\boldsymbol{\phi}^E)' = [\frac{d}{dt}(\boldsymbol{\phi}_1^E)^T, \ldots, \frac{d}{dt}(\boldsymbol{\phi}_P^E)^T]^T$.

The semidiscrete problem rewritten as a first order ordinary differential is

$$\frac{d}{dt}\left[\begin{array}{c} \mathbf{u} \\ \boldsymbol{\phi}^W \\ \boldsymbol{\phi}^E \\ \hline \mathbf{u}' \\ (\boldsymbol{\phi}^W)' \\ (\boldsymbol{\phi}^E)' \end{array}\right] = \mathcal{M}\left[\begin{array}{c} \mathbf{u} \\ \boldsymbol{\phi}^W \\ \boldsymbol{\phi}^E \\ \hline \mathbf{u}' \\ (\boldsymbol{\phi}^W)' \\ (\boldsymbol{\phi}^E)' \end{array}\right] = \left[\begin{array}{c|c} \mathcal{M}_{11} & \mathcal{M}_{12} \\ \hline \mathcal{M}_{21} & 0 \end{array}\right]\left[\begin{array}{c} \mathbf{u} \\ \boldsymbol{\phi}^W \\ \boldsymbol{\phi}^E \\ \hline \mathbf{u}' \\ (\boldsymbol{\phi}^W)' \\ (\boldsymbol{\phi}^E)' \end{array}\right]. \tag{4}$$

We propose one splitting with two steps in such way the matrix of each intermediate problem

 ISBN: 978-84-608-6082-2

has a simpler exponential. The step 1 is

$$\frac{d}{dt}\left[\begin{array}{c}\mathbf{u}\\ \boldsymbol{\phi}^W\\ \hline \boldsymbol{\phi}^E\\ \mathbf{u}'\\ (\boldsymbol{\phi}^W)'\\ (\boldsymbol{\phi}^E)'\end{array}\right] = \mathcal{M}\left[\begin{array}{c}\mathbf{u}\\ \boldsymbol{\phi}^W\\ \hline \boldsymbol{\phi}^E\\ \mathbf{u}'\\ (\boldsymbol{\phi}^W)'\\ (\boldsymbol{\phi}^E)'\end{array}\right] = \left[\begin{array}{c|c}\mathcal{M}_{11} & \mathcal{M}_{12}\\ \hline 0 & 0\end{array}\right]\left[\begin{array}{c}\mathbf{u}\\ \boldsymbol{\phi}^W\\ \hline \boldsymbol{\phi}^E\\ \mathbf{u}'\\ (\boldsymbol{\phi}^W)'\\ (\boldsymbol{\phi}^E)'\end{array}\right],$$

and the step 2

$$\frac{d}{dt}\left[\begin{array}{c}\mathbf{u}\\ \boldsymbol{\phi}^W\\ \hline \boldsymbol{\phi}^E\\ \mathbf{u}'\\ (\boldsymbol{\phi}^W)'\\ (\boldsymbol{\phi}^E)'\end{array}\right] = \mathcal{M}\left[\begin{array}{c}\mathbf{u}\\ \boldsymbol{\phi}^W\\ \hline \boldsymbol{\phi}^E\\ \mathbf{u}'\\ (\boldsymbol{\phi}^W)'\\ (\boldsymbol{\phi}^E)'\end{array}\right] = \left[\begin{array}{c|c}0 & 0\\ \hline \mathcal{M}_{21} & 0\end{array}\right]\left[\begin{array}{c}\mathbf{u}\\ \boldsymbol{\phi}^W\\ \hline \boldsymbol{\phi}^E\\ \mathbf{u}'\\ (\boldsymbol{\phi}^W)'\\ (\boldsymbol{\phi}^E)'\end{array}\right].$$

Once we have chosen the steps and we have solved exactly each step, there is still missing combining these solutions to obtain an approximation of the solution of (4). If $\psi_k^{[1]}$ and $\psi_k^{[2]}$ are the flows, with time size k, for step 1 and 2 respectively, first, we consider the second order Strang splitting

$$\mathcal{S}_k^{[2]} = \psi_{k/2}^{[1]} \circ \psi_k^{[2]} \circ \psi_{k/2}^{[1]}.$$

Then, we obtain the fourth order integrator $\mathcal{S}^{[4]}$ by composition of $\mathcal{S}^{[2]}$.

$$\mathcal{S}_k^{[4]} = \mathcal{S}_{\alpha k}^{[2]} \circ \mathcal{S}_{\beta k}^{[2]} \circ \mathcal{S}_{\alpha k}^{[2]}, \quad \text{with} \quad \alpha = \frac{1}{2-2^{1/3}}, \quad \beta = 1-2\alpha. \tag{5}$$

We remark that it is possible to save some computational cost in (5) by join together the last step in the composition of $\mathcal{S}_{\alpha k}^{[2]}$ and the first one in $\mathcal{S}_{\beta k}^{[2]}$ and similarly, the last one in the composition of $\mathcal{S}_{\beta k}^{[2]}$ and the first one in $\mathcal{S}_{\alpha k}^{[2]}$. That is,

$$\begin{aligned}\mathcal{S}_k^{[4]} &= \psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{\alpha k/2}^{[1]} \circ \psi_{\beta k/2}^{[1]} \circ \psi_{\beta k}^{[2]} \circ \psi_{\beta k/2}^{[1]} \circ \psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{\alpha k/2}^{[1]},\\ &= \psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{(\alpha+\beta)k/2}^{[1]} \circ \psi_{\beta k}^{[2]} \circ \psi_{(\alpha+\beta)k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{\alpha k/2}^{[1]}.\end{aligned}$$

4 Numerical experiments

We consider the problem with initial conditions

$$u_0(x,y) = \begin{cases} \dfrac{(x+0.2)^3(0.2-x)^3(y+0.2)^3(0.2-y)^3}{(0.2)^{12}}, & -0.2 < x, y < 0.2,\\ 0, & \text{otherwise,}\end{cases}$$

 ISBN: 978-84-608-6082-2

and $v_0(x, y) = 0$, with compact support contained in the computational domain $[-1/4, 1/4]\times[-1/4, 1/4]$. The polynomial in u_0 is chosen so that $u_0 \in C^1([-1/4, 1/4] \times [-1/4, 1/4])$.

We set $c = 1$, $s^2 = 1$ and final time $T = 4$. We study the efficiency of the splitting scheme by comparing with the fourth-order four-stage Runge-Kutta method.

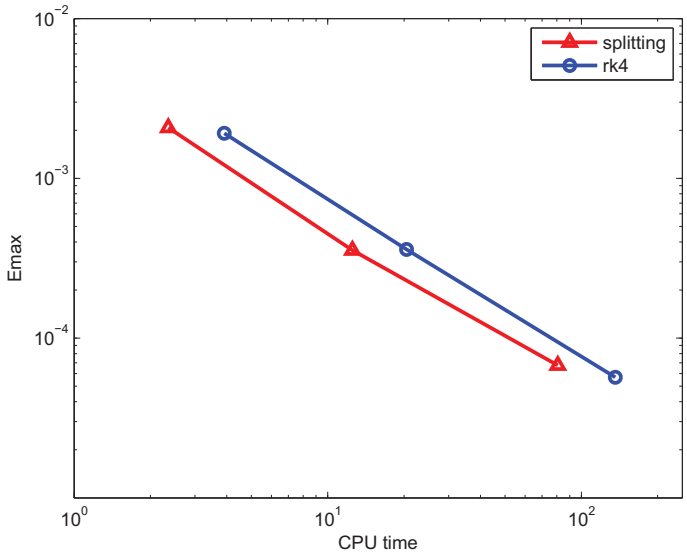


Figure 1: Emax versus CPU time for the exponential splitting integrator and the fourth-order four-stage Runge-Kutta method.

Figure 1 displays the maximum error versus CPU time for the exponential splitting integrator and the fourth-order four-stage Runge-Kutta method. It can be seen that, for the same level of accuracy, the splitting method is less costly than the Runge-Kutta method.

Acknowledgements

This work has obtained financial support from project MTM2011-23417 of Ministerio de Economía y Competitividad.

References

[1] I. Alonso-Mallo, A. M. Portillo, *Time exponential splitting technique for the Klein-Gordon equation with Hagstrom-Warburton high-order absorbing boundary conditions*, J. Comput. Phys. **311** (2016) 196–212.

[2] S. Blanes, F. Casas, A. Murua, *Splitting and composition methods in the numerical integration of differential equations*, Bol. Soc. Esp. Mat. Apl. **45** (2008), 89–145.

[3] T. Hagstrom, A. Mar-Or, D. Givoli, *High-order local absorbing conditions for the wave equation: extensions and improvements*, J. Comput. Phys. **227** (2008), 3322–3357.

[4] R.I. McLachlan, R. Quispel, *Splitting methods*, Acta Numerica **11** (2002), 341–434.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Time Series on Functional Service Life of Buildings using Fuzzy Delphi Method

A. J. Prieto[1], María-José Chávez[2], María A. Garrido-Vizuete[2], J. M. Macías-Bernal[1] and Daniel Cagigas-Muñiz[3]

[1] *Department of Architectural Construction II, University of Seville*

[2] *Departamento de Matemática Aplicada I, Universidad de Sevilla, Spain*

[3] *Department of Architecture and Computer Technology, University of Seville*

emails: andpriiba@alum.us.es, mjchavez@us.es, vizuete@us.es, jmmacias@us.es, dcagigas@us.es

Abstract

The functional service life of heritage buildings, defined as the time period during which the building fulfils the requirements for which it was designed, is a complex system that has still not been fully resolved and continues to be the object of research regarding its social, economic and cultural importance. This paper presents an application for analysing time series that reflect the state of building performance over time. To this end, historical time records are used that provided data that could be interpreted by experts in the field. The latter can then evaluate the input variables (vulnerability and risk) using the expert system for predicting the service life of buildings, Fuzzy Building Service Life (FBSL), this methodology put together the fuzzy logic tools and Delphi method. This model provides output data on the state of functionality or performance of each buildings at each moment in time whenever information records are available. The Delphi Method is used to eliminate expert subjectivity, establishing an FDM-type assessment methodology that effectively quantifies the service life of buildings over time. The application is able to provide significant data when generating future preventive maintenance programmes in architectural-cultural heritage buildings. It can also be used to optimise the resources invested in the conservation of heritage buildings. In order to validate this system, the FDM methodology is applied to some specific building examples.

Key words: Functional service life, Fuzzy Delphi Method (FDM), Expert system, Preventive conservation, Architectural heritage, Time series analysis

1 Introduction

The current economic situation has meant that resources intended for building maintenance programmes are fairly limited. As a result, there has been an increase in the number of studies focusing on how to prioritize the need for maintenance work during the service life of buildings. The aim of such studies is to estimate the moment when preventive maintenance actions or programmes need to be performed in buildings, since this type of work can considerably minimise the economic costs of conservation over time. According to the ISO 15686: 2011 international standard [18], service life can be defined as the period of time during which the building and its constituent parts fulfil the requirements for which they were designed, considered as a complex system composed of several sets of *interconnected* variables and whose links create additional information as a result of interactions. Haagenrud (2004) [2] described the number of agents causing deterioration (vulnerability and risk) that have direct and indirect consequences in terms of building maintenance and repair costs. On the other hand, the definition of the end of service life is a subjective concept that depends on criteria that may change over time.

In 2012, Macías-Bernal, Calama and Chávez presented the FBSL (Fuzzy Building Service Life) expert system [6] based on the theory of fuzzy sets [14] and intended for the diagnosis of building service life, the aim being to make predictions based on the concepts of the inherent vulnerability of buildings and external risks, where building performance in functionality terms is the output variable. This model, which will be explained briefly at the end of this section, is able to prioritize preventive conservation actions in groups of buildings with homogeneous architectural features. The FBSL system has recently been declared as compliant with the ISO 31000 : 2009 international standard [19] and the European standard EN 31010 : 2011 [20], both regulatory risk management standards and identified by the Institute of Cultural Heritage of Spain (IPCE) as useful tools for designing preventive conservation plans and risk management in heritage buildings [11]. The FBSL system has also been tested and correlated with another international model used to measure the physical degradation of materials [16] [17].

The traditional Delphi method developed by Dalkey and Helmer [1] (1963) has been widely used to obtain a steady stream of responses based on the results obtained through questionnaires.

It is one prospective method that seeks to obtain the consensus of a panel of experts based on the analysis and reflection of a well-defined problem [3]. It has three main features: anonymous reply, iteration and feedback controlled according to the statistical analysis of a group response.

It requires a long period for implementation since repeat surveys must be conducted in order to eliminate expert subjectivity and obtain uniform conclusions. On the other hand, and depending on the properties studied, it is common for the judgement of experts not to be adequately reflected in quantitative terms.

This is the case of the variables that affect the service life of a building; whenever there is certain ambiguity in their interpretation, it is more desirable for the panel of experts to reflect their preferences using linguistic terms such as "good", "very good", "average", "bad" or "very bad". These linguistic labels will subsequently be processed by fuzzy logic.

This combination between the theory of fuzzy sets and the Delphi method was proposed by Murray, Pipino and Gigch in 1985 [8] and was called the Fuzzy Delphi Method (FDM) [4].

The most important contribution of this study is the application of the FDM methodology for the analysis of historical-time records of buildings by a panel of experts using the FBSL model.

The Fuzzy Buildings Service Life (FBSL) expert system to be identified can be represented as a multi-input non-linear model: $\hat{y} = f(v(x))$, where $v(x)$ is a vector obtained from input data. In this case, the vector for each building x is the input data of the process:

$$v(x) = [v_1(x), v_2(x), \ldots, v_n(x)]$$

where $\{v_i, i = 1, \ldots, n\}$ are input variables. For the estimation of the service life to building x, the model can be represented by:

$$\hat{y}(x) = f(v(x))$$

The FBSL system has been developed following the steps established by Xfuzzy 3.0. [9] in which it is implemented: linguistic variables (input variables and output variables), rule bases and the hierarchical structure that makes up the system. To define the input parameters - specifically factors of vulnerability, static-structural, atmospheric and anthropic risk -, the following documents were reviewed: National Cathedral Plan; Law on Construction Planning: Rehabimed Method; Heritage Conservation Network; Spanish Technical Building Code; UNE 41805 : 2009 IN; ISO 15686 [6]. As a result, a total of 17 input factors (vulnerabilities and risk factors) were validated and ranked. With the collaboration of 15 professional experts in maintenance and building preservation, these factors were validated. A Delphi methodology using Opina software the property of the University of Seville was used to obtain all the experts' answers. The input variables are fuzzified in membership functions $\mu_A(v)$, in which a fuzzy set can take any value in the range of [0,1].

$$\mu_A(v) : \mathrm{U} \to [0, 1]$$

Gaussian-type membership functions are generally used, as they are considered the most appropriate, reaching a non-zero values at all points. This occurs in all the membership functions of the FBSL fuzzy inference model, except in the membership function of input variable v_1 - Geological location, whose applied membership function is trapezoidal

(established for four types of terrain). The fuzzy inference system uses the fuzzy operator "and" as a connector, which is defined as an intersection. Thus, given two sets A and B, defined on their respective universes of discourse U, the intersection of both sets is a fuzzy set $A \wedge B$, whose membership function is defined in equation (1):

$$\mu_{A \wedge B}(v_i, v_j) = T(\mu_A(v_i), \mu_B(v_j)) \tag{1}$$

where $T(x, y) = min(x, y)$ is a T-norm [13]. The fuzzy BSL system uses the minimum as connective [7].

It is well known that the core of a fuzzy system is the knowledge base comprised of two components: the data base and the rule base. The data base contains the definitions of the linguistic labels, i.e. the membership functions for the fuzzy sets. The rule base is a collection of fuzzy control rules, comprising linguistic labels, representing the expert knowledge of the controlled system. The fuzzy logic inference model, known as a generalized modus ponens, is established in the FBSL model, Equation (2), together with its hierarchical structure. In the composition of fuzzy propositions, the min-max or Mamdani inference mechanism is used [7]. This type of method works with the minimum operator as the implication function and the maximum as the aggregation operator [12]:

$$R(j): \text{ IF } v_1 \text{ is } A_1^j \text{ AND } v_2 \text{ is } A_2^j \ldots v_n \text{ is } A_n^j \text{ THEN } y \text{ is } B^1 \tag{2}$$

where $v_i(x)$ are the input (output) linguistic variables, $A_i^j(B)$ are the linguistic labels used in the input (output) variables, n is the inputs numbers and j rules numbers. The defuzzification method (the mechanism that allows the significant value discreetly representing a fuzzy set to be obtained) used by the FBSL system is the one from the Centre of the area of fuzzy set B, also known as the Centre of Gravity or Centroid [5]; it uses the centre of the area of fuzzy set B as a proxy value, $\hat{y}$. Its discrete version, which can be interpreted as a Riemann sum.

$$\hat{y} = \frac{\sum_i v_i \cdot \mu_B(v_i)}{\sum_i \mu_B(v_i)}$$

The positive properties of this method are, most notably, its continuous nature (a small change in the inputs does not imply an abrupt change in the outputs) and its non-ambiguous nature (it obtains a single value as a result of the process).

2 Delphi Fuzzy Methodology for the historical-time analysis of buildings

It is well known that modelled time series have been designed to develop an effective methodology that conforms to reality and is easy to interpret. These models are considered to be very useful applications in many scientific fields (industrial engineering, business, economic activities, etc.).

The collection of historical time series records is essential when optimising maintenance actions in buildings. The historical-time series of a building is formed by sets of data stored at different moments in time; each of these moments may be formed by one or more data records. Indeed, sometimes there are "windows" in the time series in which a single record does not give a clear and accurate idea of the state of conservation of the building; in these situations, all the accumulated data would constitute a single moment. Prieto, Macías-Bernal and Chávez (2015a) [10] took a first step in this direction by analysing the functionality of buildings through historical records. As a result, milestones were identified that significantly reduced the conservation status of the buildings studied.

Each of the professionals (i_k) belonging to the panel of experts entrusted with interpreting the different historical-time moments (y_h) of each building (x_l) in the round (j_q) will value the input variables of the FBSL: five variables of vulnerability (v_1-Geological location, v_2-Roof design, v_3-Environmental conditions, v_4-Constructive system, v_5-Preservation); 12 risk (r_6-Load state modification, r_7-Dead and live loads, r_8-Ventilation, r_9-Facilities, r_{10}-Fire, r_{11}-Inner environment, r_{12}-Rainfall, r_{13}-Temperature, r_{14}-Population growth, r_{15}-Heritage value, r_{16}-Furniture value, r_{17}-Occupancy), obtaining a Functionality Index generated by the fuzzy expert system. See Equation 3, where j_q is each round in the Delphi methodology.

Each variable involved in the historical-time application of the $FBSL_{i,j}(v(x,y))$ is described below:

$$\hat{y}(x_l, y_h) = FBSL_{i_k,j_q}(v_1(x_l, y_h)), \ldots, v_5(x_l, y_h), r_6(x_l, y_h) \ldots, r_{17}(x_l, y_h)) \tag{3}$$

- **Set of buildings** $\{x_l, l = 1, \ldots, n_1\}$

  The sample of case studies selected must be a set of buildings with uniform construction characteristics to which the 17 vulnerability and risk variables of the FBSL functionality model can be adapted. The validity of the expert system was compared dividing the buildings into two groups [6] [11] [15].

- **Moment** $\{y_h, h = 1, \ldots, n_2\}$

  The moments are made up of one or more data records. The data records may have very different characteristics and include historical pictures, paintings, engravings, construction reports, budgets, records of events, records in text format. Historical

data may contain many unique characteristics. For this reason, since the information is primarily qualitative, it is conditioned by great subjectivity when interpreted by expert professionals. Note that the methodology requires a minimum number of time points to be efficient.

- **Panel of experts** $\{y_k, k = 1, \ldots, n_3\}$

  For the experts to be able to carry out the DFM methodology, they must not know each other and there must never be a possibility for them to interact. They undertake to take responsibility for making judgements and opinions, which are the cornerstone of the method. Their profiles must cover different areas of knowledge related to the field of construction, including architecture, heritage conservation and building surveys.

  The number of experts also depends on the budget available for each study. It is generally considered that the number of experts should not be less than 7 and not more than around 30.

- **Round** $\{j_q, q = 1, \ldots, n_4\}$

  The experts are responsible for interpreting the data over the historical time series by iterating questions and answers in each *rondas* (j_q) on which the FDM method is based. A *process coordinator* group receives the responses generated in each stage. As the rounds are completed, the degree of reliability of the answers provided by the experts increases, thus generating a base of increasingly objective and reliable knowledge. As many iterations as necessary are performed among the experts to obtain sufficiently objective conclusions.

  After the action in the first round, the coordinating group calculates the appropriate statistical centralisation and dispersion parameters to observe those information records for which a fuzzier value from the experts is obtained, resulting in the drafting of the questionnaire for the next round.

3 Results and discussion

To illustrate the use of our methodology, we considered the following 20 heritage buildings located in the province of Seville (Spain).

The buildings were religious buildings with heritage features built between the 15th-16th centuries. They had other uniform political, cultural and social features. However, the chronology and stylistic characteristics of the Mudéjar buildings in the province of Seville made every building unique. Each of these buildings is located in the urban area of the corresponding locality and none are in a state of ruin or neglect [6].

A total of around 400 data records exist for the period from 1400 to the present. These include prints, paintings, photographs, records of information from newspaper archives,

manuscripts from parish archives dating from different periods describing interventions, restoration work or even possible consolidation work in the buildings, as well works certificates in the case of more recent buildings.

A group of 10 professional experts with profiles relating to the fields of Chemical Sciences, Architecture, Construction, Environment, Restoration and History were entrusted with assessing the 17 input variables of the FBSL by interpreting each of the records stored at each moment of time. Figure 1 shows the results obtained by one expert in the first round.

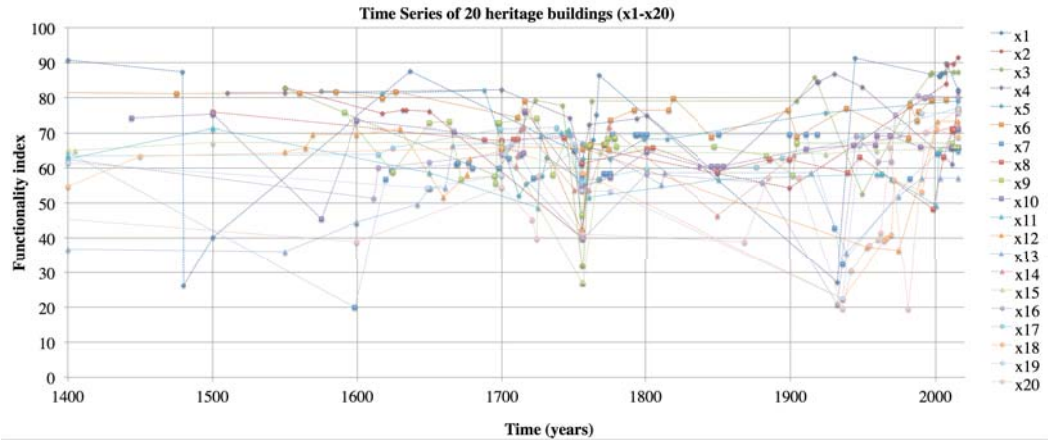


Figure 1: Historical evolution of the functionality of the uniform set of heritage buildings located in the province of Seville between 1400-2016.

As a specific case, the San Pablo Parish Church in Aznalcázar was chosen for analysis and individual representation.

In the time analysis of the San Pablo de Aznalcázar Parish Church (x_1), a total of 22 historical records dating from between 1400 and 2016 were recovered. Data collection was difficult. The historical time series data in this study were gathered manually from the parish archives owned by the Archdiocese of Seville. It was also essential to analyse the photographs that were recovered (University of Seville photographic library) as they reveal reliable information on the functional state of the building, and can also be easily compared with qualitative records in text format.

After analysing this information (see Figure 2), three unique events were identified as having had a significant influence on the functional level of the building: the first fire in the building (1480), the Lisbon earthquake (1755) and the second fire in the building four years before the Spanish Civil War (1932).

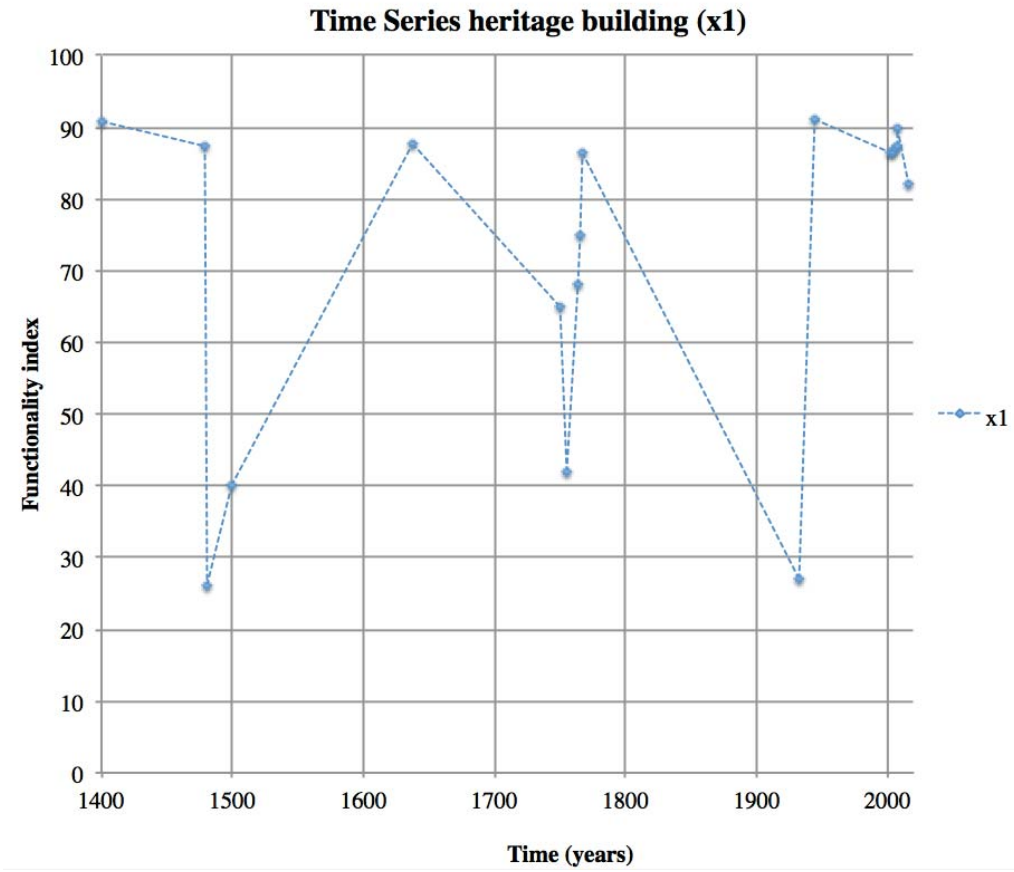


Figure 2: Historical evolution of the functionality of San Pablo de Aznalcázar parish church from 1400 to 2016.

4 Conclusions

The overall importance of sustainable development requires appropriate decisions to be taken to guarantee the service life of buildings. To achieve this, it is necessary to establish tools that can be used to define conservation and preventive maintenance plans and enhance building performance.

A new FDM-based methodology for the prediction of the service life of buildings over time by means of the analysis of historical time series is presented. This model requires records of information gathered to define the historical moments in the best possible way, and therefore achieve a better definition of building functionality. Moreover, the system is also able to effectively identify significant milestones that have compromised the life of the buildings over time.

The knowledge gathered in this study can be used to develop new methodologies based on the historical-temporal information stored and to support the taking of decisions regarding the best time to perform maintenance work, as well as to limit maintenance costs.

In terms of work in progress, FDM is currently being used in homogeneous sets of heritage buildings in southern Europe, Spain and Portugal.

This study can be extended to other buildings and other construction elements located in other regions of Europe. However, in order to carry out this application, the model must be adapted to the actual characteristics and circumstances of each location. In this sense, the analysis of the sensitivity of the model may be useful for defining and adjusting the different input variables that influence the system in order to improve the desired results and conclusions.

Acknowledgements

The authors gratefully acknowledge the support of Ministerio de Economía y Competitividad de España, Project ART-RISK - BIA2015-64878-R, Project MTM 2015-65397-P.

References

[1] N. Dalkey and O. Helmer, *An experimental application of the Delphi method to the use of experts*, Management Science **9** (1963) 458–467.

[2] S. E. Haagenrud, *Factors causing degradation. Guide and bibliography to service life and durability research for buildings and components, Joint CIB W80/RILEM TC 140*, Prediction Service Life Building Materials Components **2** (1-2) (2004) 105.

[3] C. L. Hwang and M. J. Lin, *Group decision making under multiple criteria: Methods and applications*, Springer-Verlag, 1987.

[4] A. Ishikawa, M. Amagasa, T. Shiga, G. Tomizawa, R. Tatsuta, and H. Mieno, *The max-min Delphi method and fuzzy Delphi method via fuzzy integration*, Fuzzy Sets and Systems **55** (1993) 241–253.

[5] R. Jager, H. B. Verbrugen and P. M. Bruhx, *The role of defuzzification methods in the application of fuzzy logic*, Symposium on Intelligent Components and Instruments for Control Applications (1992), Málaga, Spain.

[6] J. M. Macías-Bernal, J. M. Calama and M. J. Chávez, *Modelo de predicción de la vida útil de la edificación patrimonial a partir de la lógica difusa*, Informes de la Construcción **66**(533) (2014) p.e006.

[7] E. H. Mamdani and S. Assilian, *An experiment in linguistic synthesis with a fuzzy logic controller*, International Journal of Man-Machine Studies **7**(1) (1975) 1–13, D.O.I. 10.1016/S0020-7373(75)80002-2.

[8] T. J. Murray, L. L. Pipino and J. P. Gigch, *A pilot study of fuzzy set modification of Delphi*, Human Systems Management (1985) 6–80.

[9] F. J. Moreno-Velo, I. Baturone, A. Barriga and S. Sánchez-Solano, *Automatic tuning of complex fuzzy system with Xfuzzy*, Fuzzy sets and systems **158** (2007) 2026–2038.

[10] A. J. Prieto, J. M. Macías-Bernal and M. J. Chávez, *Series Temporales de Factores Principales para la Conservación Preventiva del Patrimonio*, 5º Congreso de Patología y Rehabilitación de Edificios (2015a), Porto, Portugal, ISBN: 78-972-752-177-7.

[11] A. J. Prieto, J. M. Macías-Bernal, M. J. Chávez and F. J. Alejandre, *Expert system for predicting buildings service life under ISO* 31000 *standard. Application in architectural heritage*, Journal of Cultural Heritage **18** (2015b) 209–218.

[12] T. J. Ross, *Fuzzy logic with engineering applications*, John Wiley & Sons, 2010.

[13] S. Weber, *A general concept of fuzzy connectives, negations and implications based on t-norms*, Fuzzy Sets and Systems **11** (1983) 115–134.

[14] L. A. Zadeh, *Fuzzy sets*, Information and Computation **8**(3) (1965) 338–353.

[15] A. J. Prieto, A. Silva, J. de Brito and F. J. Alejandre, *Functional and physical service life of natural stone claddings*, Journal of Materials in Civil Engineering **Article In Press** (2016)

[16] P.L̇. Gaspar and J. de Brito, *Service life estimation of cement-rendered facades*, Building Research and Information **36**(1) (2008) 44–55.

[17] J. L. Dias, A. Silva, C. Chai, P. L. Gaspar and J. de Brito, *Neural networks applied to service life prediction of exterior painted surfaces*, Building Research and Information **42**(3) (2014) 371–380.

[18] ISO, *Building Construction - Service Life Planning - Part 4: Service Life Planning using Building Information Modelling, ISO 15686-4, 2014, pp. 2014*,

[19] ISO, *Risk Management - Principles and Guidelines, ISO 31000, 2009, pp. 2009.*

[20] UNE EN ISO/IEC 31010, *Risk Management, Risk Assessment Techniques Focuses On Risk Assessment. Risk Assessment Helps Decision Makers Under- stand The Risks That Could Affect The Achievement Of Objectives As Well As The Adequacy Of The Controls Already In Place, 2011.*

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Extension of Newton's method for solving systems of equations when the classical Newton method fails

Higinio Ramos[1] and M. T. T. Monteiro[2]

[1] *Escuela Politécnica Superior de Zamora. Universidad de Salamanca, Avda. de Requejo, 22, 49020 Zamora. Spain*

[1] *Grupo de Computación Científica, Universidad de Salamanca*

[2] *Algoritmi R&D Center, Department of Production and Systems, University of Minho, Campus de Gualtar, 4710–057 Braga, Portugal*

emails: higra@usal.es, tm@dps.uminho.pt

Abstract

In this paper we extend the strategy presented in the article *The application of Newton's method in vector form for solving nonlinear scalar equations where the classical Newton method fails* (Journal of Computational and Applied Mathematics 275 (2015) 228–237) [1] to get the approximate solutions of a system of equations through solving an associated system obtained by the application of Newton's method. In this way, the solutions of the associated system that are not 2-cycles provide solutions of the given sytem. In most cases, the associated system can not be solved exactly. For particular starting values, solving the associated system with the Newton's method results more efficient than the application of the Newton's method to the original system. Some examples are given to illustrate the performance of the proposed strategy.

Key words: nonlinear systems, Newton's method, performance profiles

1 Introduction.

Finding the roots of a nonlinear equation $f(x) = 0$ is a classical and important problem in science and engineering. There are very few functions for which the roots can be expressed explicitly in closed form. Thus, the solutions must be obtained approximately, relying on

 ISBN: 978-84-608-6082-2

numerical techniques based on iterative processes [2]. Given an initial guess for the root, x_0, successive approximations are obtained by means of an iteration function (IF) $\Phi : X \longrightarrow X$

$$x_{n+1} = \Phi(x_n)\,, \quad n = 0, 1, 2 \ldots$$

which will often converge to a root α of the equation, provided that some convergence criterion is satisfied.

Among the iteration methods, the Newton method is probably the best known and most reliable and most used algorithm [3],[4]. There exist different results about semilocal convergence of Newton's method in which precise bounds are given for balls of convergence and uniqueness. Nevertheless, the dynamics of the Newton IF, named as N_f, may be very complicated, even for apparently simple functions and some pathological situations may occur in which Newton's method fails.

In the article by Ramos and Vigo-Aguiar [1], an alternative approach for solving pathological cases for scalar equations was presented. The strategy relies in the fact that the iteration function N_f does not have extraneous fixed points. The results for scalar functions may be extended to vector functions, and thus a similar approach may be used for solving such systems.

In this paper we present several examples of systems of equations in order to compare the performance of the application of the classical Newton method to the given system and to the associated one.

2 Main result.

The extension of the main result in [1] for systems reads

Proposition 1 *Let be $F(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^n$ a differentiable function and $N_F(\boldsymbol{x})$ the corresponding Newton IF. Any solution of the system $\{N_F(\boldsymbol{x}) = \boldsymbol{y}, N_F(\boldsymbol{y}) = \boldsymbol{x}\}$ is of one of the following types*

- *$\boldsymbol{x} = x_1, \boldsymbol{y} = x_2$ with $x_1 \neq x_2$, which means that the set $\{x_1, x_2\}$ is a 2-cycle of $N_F(\boldsymbol{x})$.*
- *$\boldsymbol{x} = \boldsymbol{y} = \alpha$, which means that α is a root of the system $F(\boldsymbol{x}) = 0$.*

The proof of this result is similar to that in [1], so the reader is referred to this reference.

3 A simple illustrative example.

Let us consider the system given by

$$\begin{cases} y^2 - 2x - 1 = 0 \\ x - y(1 + y) = 0 \end{cases}$$

which has only the solution $(0, -1)$. The plot of the two plane algebraic curves that form the system consists in two parabolas with horizontal axis that intersect tangentially on the above point.

Solving the above system with the Newton method using *Mathematica* 8.0 taking as initial guess $(x_0, y_0) = (-1.3, 1.4)$ results in a solution given by

$$x = -1.033582629 \times 10^{-25},\ y = -0.999999999999$$

after 88 iterations and CPU time of 0.0625 s., noting that the convergence is slow.

Now, being $F(x, y) = (y^2 - 2x - 1, x - y(1 + y))$, and $F'(x, y)$ the Jacobian matrix of F, after some algebraic computations we get that the Newton iteration function is given by

$$N_F(x, y) = (x, y) - (F'(x, y))^{-1} F(x, y) = (\frac{1}{2}(-1 - y), \frac{1}{2}(-1 + y))$$

and the associated system is given by

$$\{N_F(a, b) = (c, d), N_F(c, d) = (a, b)\}$$

that is, we get the system of four algebraic equations given by

$$\{1 + b + 2c = 0, b = 1 + 2d, 1 + 2a + d = 0, 1 + 2b = d\}$$

which is a linear one, and thus the Newton method gives the exact solution (except roundoff errors). This solution may be obtained easily and is $(a = 0, b = -1, c = 0, d = -1)$, and thus the solution of the initial system is $\alpha = (0, -1)$ as stated above. This is a simple example where solving the associated system results to be more efficient than solving the original system.

Note that whenever N_F is linear, the associated system will be linear too, and therefore the application of Newton's method gives an accurate result.

4 Numerical examples.

In this section we have considered different problems appeared in the literature to apply the proposed strategy. Our goal is to compare the performance of the Newton's method in solving the original system and the associated system. To do that we have considered different regions to take the starting guesses, where we have observed different behavior between both procedures. In other regions the results are similar, with the exception of the CPU time, which in some cases might be slightly larger because the associated system has twice as many equation than the original one, and most of the times is also more complicated.

 ISBN: 978-84-608-6082-2

4.1 Description of the examples considered.

4.1.1 Example 1.

Consider the system appeared in [6]

$$\begin{cases} \exp(x^2+y^2)-3=0 \\ x+y-\sin(3(x+y))=0\,. \end{cases}$$

In Figure 1 we can see the plot of the two plane curves, showing that in the rectangle $[-2,2]\times[-2,2]$ there are six roots of the system, which in fact are all the roots. These roots are the three following

$$(x_1=-1.0162459636144362, y_1=0.2566250769224934)$$
$$(x_2=-0.7411519036837556, y_2=0.7411519036837556)$$
$$(x_3=-0.2566250769224934, y_3=1.0162459636144362)$$

together with those symmetric of the above with respect to the diagonal line $y=x$. We have considered the region $([-2.3,0.1]\times[-2,-0.2])\bigcup([1.1,2.3]\times[0.2,2])$ with stepsizes $h_x=h_y=0.2$ for taking the initial guesses to test the performance of the two approaches.

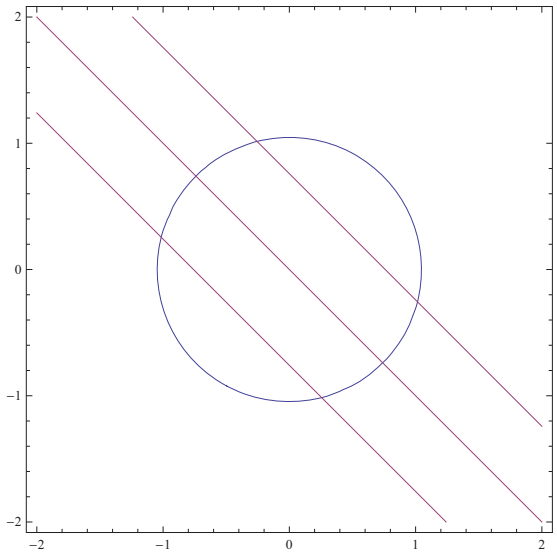


Figure 1: Plot of the curves in Example 1.

4.1.2 Example 2.

Consider the system in [7] given by he two equations

$$\begin{cases} -4 - 2x + x^3 - 3xy^2 = 0 \\ -2y + 3x^2y - y^3 = 0 \end{cases}$$

where each equation corresponds respectively to the real and imaginary parts of the equation $z^3 - 2z - 4 = 0$, with $z \in \mathbb{C}$. The system has three roots, $(2, 0), (-1, 1), (-1, -1)$. In Figure 2 the plot of the two curves with the roots can be seen. We have considered the region $([-0.8, 0.8] \times [-0.8, 0.8]) \bigcup ([-4.3, -2.3] \times [-0.8, 0.8])$ with stepsizes $h_x = h_y = 0.05$ for taking the initial guesses to test the performance of the two approaches.

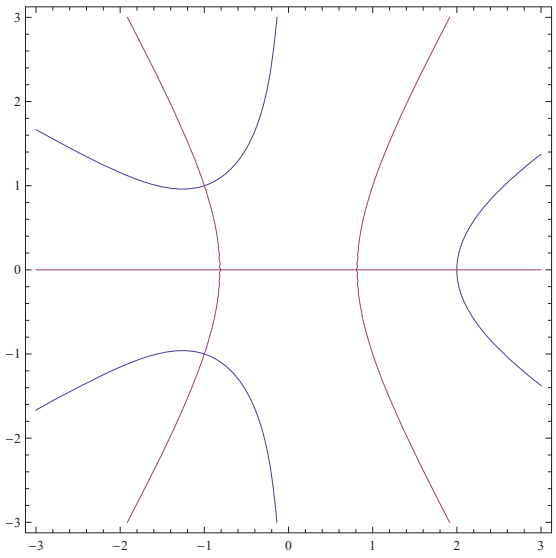


Figure 2: Plot of the curves in Example 2.

4.1.3 Example 3.

Consider the following system which is a modification of that in [5] p.119

$$\begin{cases} x^2 + y - 3 = 0 \\ x + y^3/8 - 1 = 0\,. \end{cases}$$

This system is formed by two polynomial equations, and has two roots

$$(x_1 = -0.7638749087120703, y_1 = 2.4164951238401262)$$
$$(x_2 = 2.2732612217399874, y_2 = -2.1677165822667804)$$

 ISBN: 978-84-608-6082-2

as can be seen in Figure 3. Doing as in [5], if the starting point in the Newton method is taken as $(x_0 = 1, y_0 = -1)$ the result is not correct, resulting after 182 iterations in $x = 1.484654, y = 0.947670$, and thus the Newton method fails. Nevertheless, solving the associated system with the Newton method taken as starting point $(a_0 = c_0 = 1, b_0 = d_0 = -1)$, after 23 iterations we get the first of the above roots with 50 digits of accuracy. We have considered the region $([-4, 4] \times [-4, 4])$ with stepsizes $h_x = h_y = 0.2$ for taking the initial guesses to test the performance of the two approaches.

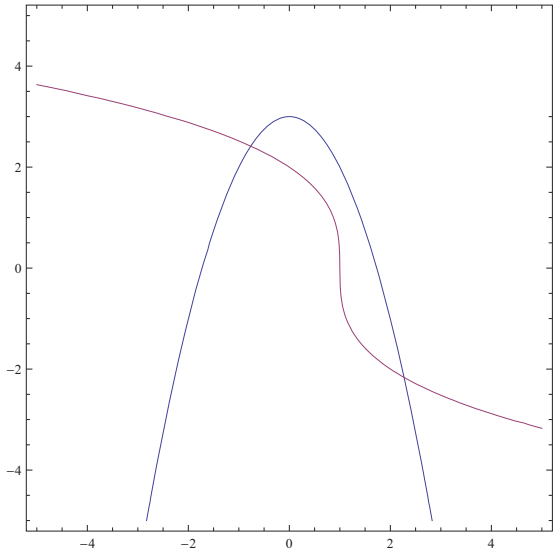


Figure 3: Plot of the curves in Example 3.

4.1.4 Example 4.

Consider the system in [8] given by the four equations

$$\begin{cases} x + 10y = 0 \\ \sqrt{5}(z - t) = 0 \\ (y - 2z)^2 = 0 \\ \sqrt{10}(x - t)^2 = 0 \end{cases}$$

which has only the root $(0, 0, 0, 0)$. In this case the associate system is linear, and thus the application of Newton method has no problem in finding the root. We have considered the region $([1.2, 1.8] \times [1.2, 1.8] \times [0.3, 0.7] \times [2, 2.4])$ with stepsizes $h_x = 0.06, h_y = 0.07, h_z = 0.07, h_t = 0.13$ for taking the initial guesses to test the performance of the two approaches.

4.1.5 Example 5.

Consider the system appeared in [9] given by

$$\begin{cases} x+y-1=0 \\ 2x+y+2z-2=0 \\ x+y+z-t=0 \\ \dfrac{y^2z}{x^2t}-0.647^2=0 \end{cases}$$

which has only the real root $(x = 0.422499, y = 0.577501, z = 0.288751, t = 1.288750)$. We have considered a region to select the initial guesses far from the root, namely, $([20.2, 20.8] \times [12.2, 12.8] \times [14.3, 14.7] \times [18.3, 18.7])$ with stepsizes $h_x = 0.06, h_y = 0.07, h_z = 0.07, h_t = 0.13$ to test the performance of the two approaches.

4.2 Performance profiles.

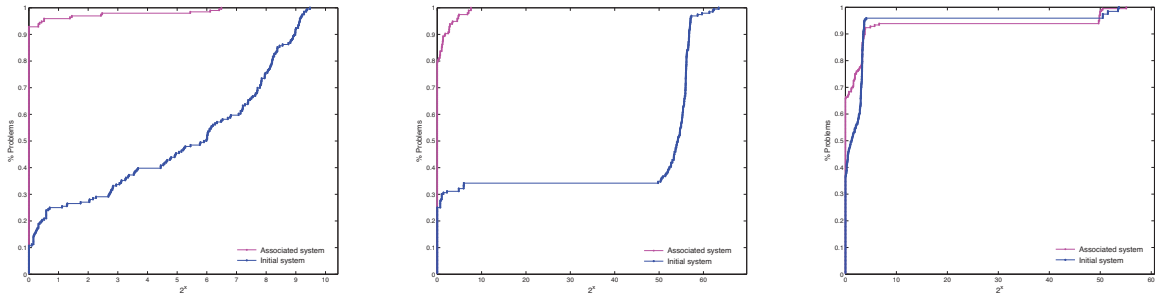
In order to analyze the relative performance of both approaches (associated system and original system), the performance profiles of Dolan and Moré [10] were used. Several runs were made for the five exemples, considering different initial approximations for both approaches. For the two-dimensional grid of starting points indicated in each example, we have considered the following performance metrics: the number of iterations, the error and the CPU time. The error is measured by the maximum absolute value of the component functions evaluated at the provided roots, that is, $\max_{1\leq i\leq n}\{|F_i(\alpha)|\}$ where α is the requested root of the system $F(\mathbf{x}) = 0$.

4.2.1 Example 1.

Figure 4 reports results for Example 1. The associated system approach outperforms the initial system approach in more than 90%, for the iterations, and more than 80% for error metrics. CPU time performances are similar.

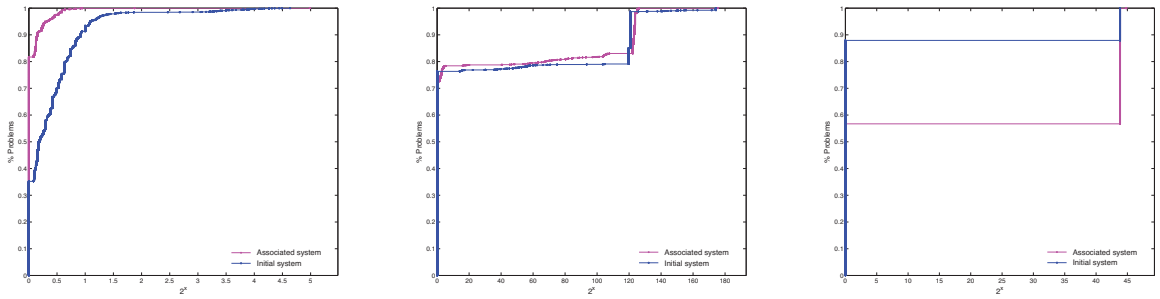
4.2.2 Example 2.

Figure 5 pictures results for Example 2. In respect of number of iterations the associated system approach performs better than the initial system more than 80% of the runs. The error metric is similar for both approaches. The initial system approach has better CPU time in more than 80% of the runs.



(a) Iterations (b) Error (c) CPU time

Figure 4: Example 1 performance profiles.

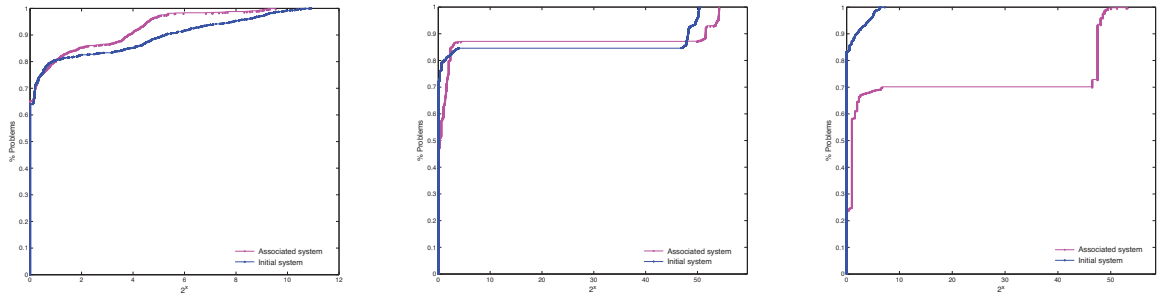


(a) Iterations (b) Error (c) CPU time

Figure 5: Example 2 performance profiles.

 ISBN: 978-84-608-6082-2

4.2.3 Example 3.

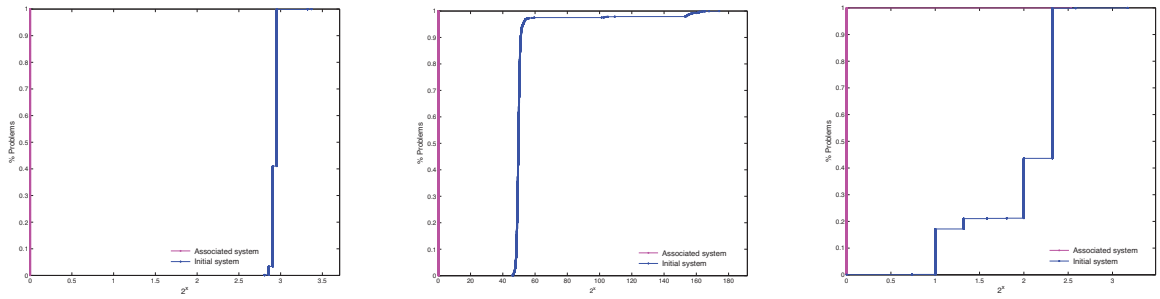


(a) Iterations (b) Error (c) CPU time

Figure 6: Example 3 performance profiles.

Figure 6 presents results for Example 3. For the number of iterations and error, the associated system approach performs better than the initial system. However, the initial system approach has better CPU time in more than 80% of the runs.

4.2.4 Example 4.

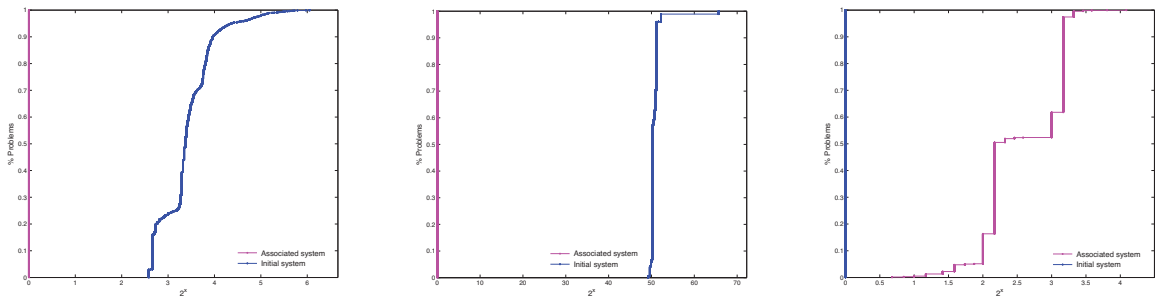


(a) Iterations (b) Error (c) CPU time

Figure 7: Example 4 performance profiles.

Figure 7 pictures results for Example 4. For all metrics, the associated system approach performs much better than the initial system. This results are due to the fact that the associated system is a linear one.

4.2.5 Example 5.



(a) Iterations (b) Error (c) CPU time

Figure 8: Example 5 performance profiles.

Figure 8 shows the results for the Example 5. Considering the number of iterations and the error, the associated system approach presents better performance than the initial system. The initial system approach has better CPU time.

5 Conclusions

The strategy presented here, consisting in solving by Newton's method the associated system to get the roots of a given system of equation, is not the ultimate one. That is, when Newton's method works well with the original system, we must use this. Only when Newton's method does not work well with the original system, we should consider to solve the associated system. We must keep in mind that the associated system has twice equations than the original system, and this generally complicates its resolution. In specific cases, namely when the associated system is linear, the proposed strategy works clearly better. The numerical experiments we have presented validate these claims.

Acknowledgements

The second author has been supported by the ALGORITMI R&D Center and project PEst-UID/CEC/00319/2013.

References

[1] H. Ramos, J. Vigo-Aguiar, *The application of Newton's method in vector form for solving nonlinear scalar equations where the classical Newton method fails*, J. of Comput. and App. Math. **275** (2015) 228–237.

[2] J. M. Ortega,W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Banach Spaces*, Academic Press, NewYork, 1970.

[3] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, SIAM, Philadelphia, 2003.

[4] P. Deuflhard, *Newton Methods for Nonlinear Problems. Affine Invari- ance and Adaptive Algorithms*, Springer, Berlin, 2011.

[5] J.L. Awange, B. Paláncz, *Geospatial algebraic computations*, Springer-Verlag, 3th Ed., Berlin, 2016.

[6] M. Amrein, T. P. Wihler, *An adaptive Newton-method based on a dynamical systems approach*, Commun. Nonlinear Sci. Numer. Simulat. **19** (2014) 2958–2973.

[7] M.Amrein and T.P.Wihler, *An adaptive Newton-method based on a dynamical systems approach*, Research Report 2013 - 09.

[8] M. J. D. Powell, *An iterative method for finding stationary values of a function of several variables*, Comput.J. **5** (1962) 147–151.

[9] R. E. Martínez Solano, *Aplicación de métodos secante-estructurados en la solución de sistemas de ecuaciones mixtos*, Trabajo de Grado, Universidad del Valle, Barranquilla, 1999.

[10] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming **91** (2002) 201-213.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Monte Carlo Approach for the Pricing of European Multi-Asset Options

A. Rasulov[1] and G. Raimova[2]

[1] *Department of Mathematical Modelling and Informatics, University of World Economy and Diplomacy*

[2] *Department of ICT, Academy of Public Administration*

emails: asrasulov@gmail.com, raimova27@gmail.com

Abstract

This work concerns to the application of Monte Carlo methods for a class of linear partial differential equations (PDEs). This PDEs we can be find in mathematical finance when calculating the price of European Multi-Asset Options. The calculation methods are based on the simulation of solutions of Ito stochastic differential equations (SDEs) and the application of Monte Carlo algorithms to the PDEs. Numerical algorithms for SDEs ([2]), which can be used as the basis of Monte Carlo simulations, are discussed, and some theoretical results of the new approaches are presented.

Key words: Monte Carlo methods, multi-asset options, stochastic differential equation, simulation, Brownian motion
MSC 2000: AMS codes (optional)

1 Introduction

This article concerns to application of Monte Carlo methods for the pricing European Multi-Asset Options. Application of Monte Carlo methods to the class of linear partial differential equations devoted the books ([1],[5]). The complexity of a Monte Carlo method is typically polynomial in the dimension, whereas the deterministic method is typically exponential. In the context of PDEs, a useful feature of Monte Carlo methods is that they allow the solution to be found at just one point, if required (with associated saving in computation), whereas deterministic methods necessarily find the solution at large number of points simultaneously. This property of Monte Carlo methods can be particularly useful in problems such as option pricing, where the value of an option is required only at the time of striking, and for the state of the market at that time.

2 Exact and Monte Carlo Solution Stochastic Models of Multi-Assets Pricing

In order to price multi-asset options, we firstly need to establish the price movement model for the underlying multi-assets. Let X_i be the price of the i -th risky asset $(i = 1, ..., m)$. In general the prices of multi-assets can be modelled as:

$$\frac{dX_i}{X_i} = \mu_i dt + \sum_{j=1}^{m} \sigma_{ij} dW_j, \ (i = 1, ..., m), \tag{1}$$

where dW_i $(i = 1, ..., m)$ are one dimensional standard Brownian motions $E(dW_i) = 0$, $Var(dW_i) = dt$, here $Cov(dW_i, dW_j) = 0$, $(i \neq j)$, $Cov(.,.)$ denotes the covariance, μ_i - is expected return rate of X_i, σ_{ij} -is a component of the instantaneous standard deviation of the rate provided by X_j, which may be attributed to the X_i. SDE (1) can be written in the vector form $d\vec{X} = \vec{a}dt + [\sigma]\, d\vec{W}_t$, where

$$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}, \vec{a} = \begin{bmatrix} \mu_1 X_1 \\ \vdots \\ \mu_m X_m \end{bmatrix}, \vec{W} = \begin{bmatrix} W_{1t} \\ \vdots \\ W_{mt} \end{bmatrix}, [\sigma] = [X] \ [\sigma_0],$$

$$[X] = \begin{bmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_m \end{bmatrix}, [\sigma_0] = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1m} \\ \vdots & & \vdots \\ \sigma_{m1} & \dots & \sigma_{mm} \end{bmatrix}$$

Let $(X_1, ..., X_m)$ be m risky assets (e.g., stock, foreign exchange rate,...), satisfying geometric Brownian motion (1). Let V be an option derived from underlying assets $(X_1, ..., X_m)$, as a function of $m + 1$ variables $(X_1, ..., X_m)$ and t:

$$V = V(X_1, ..., X_m, t) \tag{2}$$

Let q_i be dividend rate of asset X_i, and $a_{ij} = \sum_{k=1}^{m} \sigma_{ik}\sigma_{jk}$ $(i, j = 1, ..., m)$ i.e. $A = [a_{ij}] = \sigma_0\sigma_0^T$, where σ_0^T is the transpose of matrix σ_0. Let r is risk free interest rate. From the Ito formula for the multivariate stochastic process, it is easy to get Black-Scholes equation for multi-asset options:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sum_{i,j=1}^{m} a_{ij} X_i X_j \frac{\partial^2 V}{\partial X_i \partial X_j} + \sum_{i=1}^{m} (r - q_i) X_i \frac{\partial V}{\partial X_i} - rV = 0 \tag{3}$$

This equation is called the Black-Scholes equation for multi-asset options. Since$A = [a_{ij}]$ is a symmetrical nonnegative matrix the equation (3) a multidimensional parabolic

equation. Here r is an instantaneous (in very short term) risk-free rate of interest. If λ_i is a market price of risk of X_i than $\mu_i - r = \sum_{i=1}^{m} \lambda_i \sigma_{ij}$. Denoting the option payoff function at maturity ($t = T$) by $f(X_1, ..., X_m)$, then the mathematical model of the European multi-assets option is: solve PDE (3) in domain $D : \{0 \leq X_i < \infty \ (i = 1, ..., m); 0 \leq t \leq T\}$ with the terminal condition

$$V(X_1, ..., X_m, t) = f(X_1, ..., X_m), \ x \in R_m^+, \ t = T. \tag{4}$$

By transformation $Y_i = LnX_i$ the equation (3) becomes

$$\frac{\partial V}{\partial t} + \frac{1}{2} \sum_{i,j=1}^{m} a_{ij} \frac{\partial^2 V}{\partial y_i \partial y_j} + \sum_{i=1}^{m} (r - q_i - \frac{a_{ii}}{2}) \frac{\partial V}{\partial y_i} - rV = 0 \tag{5}$$

$$V(y_1, ..., y_m, T) = f(e^{y_1}, ..., e^{y_m}) \tag{6}$$

The terminal value problem to multidimensional equation (5)– (6) has a solution. Back to the original variables $(X_1, ..., X_m)$ we obtain the European multi-asset option pricing formula ([3]):

$$V(X, t) = \left[\frac{1}{2\pi(T-t)} \right]^{\frac{m}{2}} \frac{e^{-r(T-t)}}{|\det A|^{\frac{1}{2}}} \times$$

$$\times \int_0^\infty \cdots \int_0^\infty \frac{f(\xi_1, ...\xi_m)}{\xi_1 \cdots \xi_m} \exp \left[\frac{\vec{\alpha}^T A^{-1} \vec{\alpha}}{2(T-t)} \right] d\xi_1 \cdots d\xi_m \tag{7}$$

where

$$\vec{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}, \ \alpha_i = Ln \frac{X_i}{\xi_i} + (r - q_i - \frac{a_{ii}}{2})(T-t), \ (i = 1, ..., m). \tag{8}$$

The equation (7) is called the Black-Scholes formula for European multi-asset options.

However, this is a multiple integral with singularities in the integrand. When large numbers of assets are involved, the integral has a high multiplicity and is very difficult to evaluate. Thus, finding a closed-form expression is only the first step in solving the pricing problem of the European multi-asset option. We still need to find a simple way to get a special solution to the problem for each concrete form of the payoff function. That is a way for the solution of the multi-asset problem (5) and (6) we suppose improved Monte Carlo algorithm which proposed in [4] for the solutions of PDE (3) with the boundary conditions similarly (4). Here we recall the probabilistic representation for the solution $V(x, t)$ of the Cauchy problems (3) and (4). In fact, the solution to problems (3) and (4) has various probabilistic representations which could be calculated by Monte Carlo methods:

$$V(x, t) = E\left[V\left(X_{x,t}(T), T\right) Y_{x,t,1}\left(T\right)\right] = E\left[f\left(X_{x,t}(T)\right) Y_{x,t,1}\left(T\right)\right]$$

 ISBN: 978-84-608-6082-2

where $X_{x,t}(s), Y_{x,t,y}(s),\ s \geq t$, is the solution of problem Cauchy for the following system of stochastic differential equations:

$$\begin{cases} dX = c(X,t)ds + \sigma(X,t)dw(s), \quad X(t) = x, \\ dY = -rYds, \quad Y(t) = y, \end{cases} \tag{9}$$

$(x,t) \in Q$, where $Q = \{0 \leq x_i < \infty, \quad i = 1,m; \quad 0 \leq t < T\}$.

In (9) $w(s) = \left(w^1(s), w^2(s), \dots w^m(s)\right)^T$ - m-dimensional standard Wiener process, Y - being scalar processes, $c(x,s)$ - m- dimensional column vector, compounded from the coefficients $c_i(x,t) = (r - q_i)X_i$, $\sigma(x,s)$ -matrix with dimensions $m \times m$, where $\sigma(x,s)\ \sigma^T(x,s) = a(x,s)$, and $a(x,s) = \{a_{ij}X_iX_j\}$, $\quad i,j = 1,m$.

For the solution of to system of stochastic differential equations (9) we should apply the theory of weak solutions [4]. If we will use direct Euler methods with constant step size, the price of options could be negative due to the Wiener process. For the numerical solution of (9) we suppose the following new proposed algorithm. Let us consider the time discretization $T = t_N > t_{N-1} > \dots > t_0$ with the steps $h = \frac{T-t_0}{N}$. Let $(x_0,t_0) \in Q$, $\quad y_0 = 1$, ξ - a uniformly distributed point on the open surface $\partial S_1(0) \subset R^m$ with the unit radius with the center point of origin, $\rho_0 = \sqrt{m \cdot h}$. We introduce spheroid $S(x_0,t_0,\rho)$, which is obtained with the help of linear transformation $\rho \cdot \sigma(x_0,t_0)$ ball $S_1(0)$ with the shift $x_0 + \frac{\rho^2}{m}c(x_0,t_0)$. Let $\rho = \rho_0$. If $S(x_0,t_0,\rho) \not\subset R^m_+$, we will find $\rho < \rho_0$ such $S(x_0,t_0,\rho)$ which touch the boundary R^m_+. Let $t_1 = t_0 + h$, $x_1 = x_0 + c(x_0,t_0)\frac{\rho^2}{m} + \rho\sigma(x_0,t_0) \cdot \xi$ and $y_1 = y_0\left(1 - r\frac{\rho^2}{m}\right)$. It is clear that the point x_1 will lay on the surface of spheroid $S(x_0,t_0,\rho)$. Next points are simulated analogously. Let $x_{N+k} = x_k$ and $\Im_k$ - sigma-algebra, filtrated by $(\xi_1,\xi_2,\dots,\xi_k)$ isotropic vectors $\Im_k = \sigma(\xi_1,\xi_2,\dots,\xi_k)$. It is possible to prove that constructed sequence of coordinate $\{x_k\}_k$ is martingale.

Lemma. *Let $\Im_k$ - sigma-algebra, filtrated by $(\xi_1,\xi_2,\dots,\xi_k)$ isotropic vectors $\Im_k = \sigma(\xi_1,\xi_2,\dots,\xi_k)$. The sequence of coordinate vector process $\{x_k\}_k$ - is martingale relatively $\Im_k$.*

Proof: From the definition $\Im_k$ follow, that x_k is $\Im_k$ - measured. From the property of conditional mathematical expectations we get

$$E(x_{k+1}/\Im_k) = E\left(x_k + c(x_k,t_k)\frac{\rho^2}{m} + \rho\sigma(x_k,t_k)\ \xi_k/\Im_k\right) =$$
$$= \frac{1}{|\partial S_1(0)|}\int\limits_{\partial S_1(0)}\left(x_k + c(x_k,t_k)\frac{\rho^2}{m} + \rho\sigma(x_k,t_k)\ \omega\right)d\omega = x_k,$$

where $\partial S_1(0)$ unit surface area in R^m and $d\omega$ - element of surface. Lemma is proved.

Let $(x_0,t_0) \in Q$, $S = S(x_0,t_0,\rho)$ spheroid which defined above, P - cylinder, $P = S \times [t_0,t_0+h)$. Let as assume $t_1 = t_0 + h$, $x_1 = x_0 + c(x_0,t_0)\frac{\rho^2}{m} + \rho\sigma(x_0,t_0) \cdot \xi$ and $y_1 = y_0\left(1 - r\frac{\rho^2}{m}\right)$. The next theorem gives the order of accuracy one step approximation of (9).

Theorem 1. *Let u be restricted solution in $\overline{P}$ problem (3) and (4), with continuous derivatives $D_x^m D_t^k u$, $0 \leq m+2k \leq 4, \quad k=0,1$. Let $X_{x_0,t_0}(s)$, $Y_{x_0,t_0,y_0}(s)$ - a solution of (9) at the points $(x_0,t_0) \in Q$, τ - be a hitting time of the process $(X_{x_0,t_0}(s),\ s)$ on the boundary P (the point $(X_{x_0,t_0}(\tau),\ \tau)$ belong upper side surface of cylinder P). Than the order of accuracy (error) of restricted solution for one step $(x_0,t_0) \to (x_1,t_1) \in \overline{P}$ will be*

$$E\left[u\left(x_1,t_1\right)y_1 - u\left(X_{x_0,t_0}\left(\tau\right),\tau\right)Y_{x_0,t_0}\left(\tau\right)\right] = O\left(h^2\right) \tag{10}$$

The proof of this theorem is given in ([5]). This theorem has showed that the order of accuracy of the solution in one step approximation by discretization of stochastic process equals $O(h^2)$. Now we will construct the numerical algorithm of solution of problem (2.3) and (2.4) based on one step approximation. Let

$$t_{k+1} = t_k + h,$$

$$x_{k+1} = x_k + c(x_k,t_k)\frac{\rho_k^2}{m} + \rho_k \sigma(x_k,t_k)\cdot \xi_k,$$

$$y_{k+1} = y_k\left(1 - r\frac{\rho_k^2}{m}\right), \quad y_0 = 1.$$

Here $\xi_1, \cdots, \xi_k, \ldots$ the sequence of independent isotropic vectors in R^m.

Theorem 2. *Let V be a solution in given domain Q problem (3) and (4), with continuous derivatives $D_x^m D_t^k u$, $0 \leq m+2k \leq 4, \quad k=0,1$. The following estimated error of the solution of method based on one step approximation will be*

$$E\left[f\left(X_{x,t}(T)\right)Y_{x,t,1}(T) - V\left(x,t\right)\right] = O\left(h\right)$$

Proof: Let $(x,t) = (x_0,t_0)$. Since $V(X_1,...,X_m,t) = f(X_1,...,X_m)$

$$E\left[f\left(X_{x,t}(T)\right)Y_{x,t,1}(T) - V\left(x,t\right)\right] = E\left[V\left(X_{x,t}(t_N),t_N\right)Y_{x,t,1}(t_N) - V\left(x,t\right)\right] =$$

$$= \sum_{k=0}^{N-1} E\left[V\left(x_{k+1},t_{k+1}\right)y_{k+1} - V\left(x_k,t_k\right)y_k\right]$$

According to theorem 2, one step approximations order of accuracy is $O\left(h^2\right)$, consequently each term of sum satisfied following inequality

$$E\left[V\left(x_{k+1},t_{k+1}\right)y_{k+1} - V\left(x_k,t_k\right)y_k\right] \leq Kh^2.$$

Since $h = \frac{T-t_0}{N}$ as a result we will get

$$E\left[f\left(X_{x,t}(T)\right)Y_{x,t,1}(T) - V\left(x,t\right)\right] \leq N\cdot Kh^2 = K_1 h$$

The theorem is proved.

 ISBN: 978-84-608-6082-2

3 Conclusion

The proposed method has two type (approximation and usual classic Monte Carlo) of errors. We suppose in our further work will do several numerical experiments, compare with the results which were obtained in ([3],[4]) and will give error analysis.

References

[1] S. M. Ermakov, V. V. Nekrutkin and A. S. Sipin, *Random Processes for Classical Equations of Mathematical Phisics*, Kluwer Acad. Publ., 1989.

[2] A. Friedman, *Stochastic Differential equations and Applications*, Volume 1, Academic Press., 1975.

[3] L.Jiang, *Mathematical Modeling and Methods of Option Pricing*, World Scientific Publishing Co.Pte.Ltd, Singapore, 2005.

[4] G. N. Milstein, J. G. M. Schoenmakers, *Monte Carlo Construction of Hedging Strategies Against Multi-Asset European Claims*, Stochastic and Stochastic Reports, Vol.73, (1-2), (2002) 125–157.

[5] A. Rasulov, M. Mascagni and G. Raimova, *Monte Carlo Methods for the Solution of Linear and Nonlinear Boundary Value Problems*, Printing house of UWED, Tashkent, 2006.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On the quasi-positive systems

Beatriz Ricarte[1] **and Sergio Romero-Vivó**[1]

[1] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Spain*

emails: bearibe@mat.upv.es, sromero@imm.upv.es

Abstract

In this paper we deal with discrete-time linear control systems in which the state is constrained to lie in the positive orthant R^n_+ independently of the inputs involved, that is, the inputs can take negative values. Such (quasi-positive) systems appear for example in ecology models where the removal of individuals from a population is described. Controllability and reachability are fundamental properties of a system that show its ability to move in space, which are analyzed throughout the text, paying special attention to the single-input case.

Key words: Quasi-positive systems, Reachability, Single-Input Systems.

1 Introduction

In many biological processes for instance, metabolism and drug ingestion, the state and input variables must be nonnegative values. The need to understand the properties of these kind of systems induced the development of Positive Systems Theory, which is well documented in the bibliography [2, 3].

Nevertheless, there are many applications where it is not necessary to be so restrictive and only the state must be nonnegative. For example, in several economic models [7], or in population ecology where to describe the removal of individuals from a population it is required that the control can take negative values ([1, 4, 6]). These processes can be represented by the discrete-time linear system with n states and m inputs

$$x(t+1) = Ax(t) + Bu(t), \quad t \in \mathbb{N}_0 \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are nonnegative matrices, the state $x(\cdot) \in \mathbb{R}^n$ is a nonnegative vector and without restrictions for the control vector $u(t) \in \mathbb{R}^m$, which are said to be *quasi-positive systems* according to reference [1]. From now on, let us denote them by $(A, B)_q \geq 0$.

 ISBN: 978-84-608-6082-2

2 Basic Notions

Quasi-positive systems $(A, B)_q \geq 0$ could be considered as a non-trivial middle ground between Linear Systems and Positive Linear Systems. Only few authors have contributed to this topic. It is worth mentioning Guiver *et al.* [5] who dealt with the controllability property of these systems, which was termed *positive state controllability*. It is understood that $(A, B)_q \geq 0$ is positive state controllable in finite time if every nonnegative initial state x_0 can be transferred to any other nonnegative final state x_f in a finite number of steps and additionally it is maintained the nonnegativity of the states involved. This property is equivalent to reachability ($x_0 = 0$) and null-controllability ($x_f = 0$) with positive state, which we briefly call *quasi-positive reachability* and *quasi-positive null-controllability*, respectively (or simply, reachability and null-controllability in this context).

Guiver *et al.* have demonstrated that under certain specific assumptions the problem of positive state controllability is equivalent to positive input controllability of a related positive system. Unfortunately, there are many biological systems where those assumptions do not hold. In the following sections it is analyzed this properties for any quasi-positive system and more particulary, the single-input case and its applicability to a population ecology model.

3 Preliminary results

Although from a theoretical point of view any timing, finite or infinite, is possible to achieve reachability, the next results are always under conditions of finiteness to be useful in real-life practice.

Given a quasi-positive system $(A, B)_q \geq 0$, we have proved that if any canonical vector e_i, $i = 1, 2, \ldots, n$ is reachable (in finite time), then any nonnegative state $x \in \mathbb{R}^n_+$ is reachable, then the system is quasi-positive reachable. Moreover, this nonnegative state $x \in \mathbb{R}^n_+$ will be reached in at most n steps, where n is the dimension of $(A, B)_q \geq 0$, and additionally the reachability matrix $\mathcal{R}_n(A, B) = [B\ AB\ A^2B\ \cdots\ A^{n-1}B]$ has rank equal to n.

Example 1 *Let* $(A, b)_q \geq 0$ *with*

$$A = \begin{bmatrix} * & 1 & 0 \\ * & 0 & 1 \\ * & 0 & 1 \end{bmatrix}, \text{ and } b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \mathcal{R}_3(A, b) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Clearly, canonical vector e_3 *can be reached in one step. In the same way, canonical vector* e_2 *(*e_3*) in two steps (three steps) using the sequence of controls* $u(0) = 1$, $u(1) = -1$ *(*$u(0) = 1$,

$u(1) = -1, u(2) = 0)$ *which maintains the nonnegativity of* x *in each previous step. Therefore,* $(A, b)_q \geq 0$ *is reachable. Observe that in this case the rank of the reachability matrix* $\mathcal{R}_3(A, b)$ *is 3 and that this same system is not positive reachable (using only nonnegative inputs).*

The inverse is not true as the following example proves.

Example 2 *Let* $(A, b)_q \geq 0$ *be a system given by the pair*

$$A = \begin{bmatrix} * & 1 & 0 \\ * & 1 & 1 \\ * & 1 & 1 \end{bmatrix}, \text{ and } b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \Rightarrow \mathcal{R}_3(A, b) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 1 & 2 \end{bmatrix}.$$

Its reachability matrix $\mathcal{R}_3(A, b)$ *has rank equal to 3 but it is not quasi-positive reachable because the sequence of controls to attain* e_1 *is* $u(0) = 1$*,* $u(1) = -2$ *and* $u(2) = 0$*, which would have previously steered in two steps the initial state toward the non-positive state* $x(2) = [0\ 0\ -1]^T$*.*

4 Single-Input Quasi-Positive Systems

A generalization of the structure given by the matrices involved in example 1 provide us with a class of quasi-positive reachable systems. One can prove that a single-input quasi-positive system is reachable only if the input vector b is monomial, that is, b is a positive multiple of a canonical vector. Moreover, any pair similar by permutation to the pair $(\hat{A}, \hat{b})_q \geq 0$ where

$$\hat{A} = P^T A P = \begin{bmatrix} * & + & 0 & \cdots & 0 \\ * & 0 & + & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ * & 0 & 0 & \cdots & + \\ * & * & * & \cdots & * \end{bmatrix}, \text{ and } \hat{b} = P^T b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ + \end{bmatrix}$$

with $+$ a positive entry and $*$ a nonnegative entry is quasi-positive reachable or even at the same time quasi-positive null-controllable under certain values of the nonnegative entries, that is, quasi-positive controllable. This pattern appears for example in a Fibonacci model used to describe the evolution of a rabbit population subjected to hunting (see [3]).

In the multiple input case, a quasi-positive system $(A, B)_q \geq 0$ can be reachable having non-monomial columns in the input matrix B which are necessary to guarantee such property.

 ISBN: 978-84-608-6082-2

Example 3 *Let* $(A,b)_q \geq 0$ *be a system given by*

$$A = \begin{bmatrix} * & 1 & 0 \\ * & 1 & 0 \\ * & 1 & 0 \end{bmatrix}, \text{ and } b = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

It is clear that canonical vector e_3 *can be reached in one step. In the same way, canonical vector* e_2 *(*e_3*) in two steps (three steps) using the sequence of controls* $u(0) = 1$*,* $u(1) = -1$ *(*$u(0) = 1$*,* $u(1) = -1, u(2) = 0$*) which maintains the nonnegativity of* x *in each previous step. Therefore,* $(A,b)_q \geq 0$ *is reachable but it is not positive reachable (using only nonnegative inputs).*

Acknowledgements

This work has been partially supported by Ministerio de Ciencia e Innovación through Grant DPI-2013-46982-C2-1-R and Grant MTM-2013-43678-P.

References

[1] R. Cantó, B. Ricarte and A. M. Urbano, *Modelling big game populations when hunting is age and sex selective*, Mathematical and Computer Modelling **57** (2013) 1744–1750.

[2] P. G. Coxson and H. Shapiro, *Positive input reachability and controllability of positive systems*, Linear Algebra and its Applications **94** (1987) 35–53.

[3] L. Farina and S. Rinaldi, *Positive Linear Systems: Theory and Applications*, John Willey and Sons, Canada, 2000.

[4] J. R. Fieberg, K. W. Shertzer, P. B. Conn, K. V. Noyce and D. L. Garshelis, *Integrated population modeling of black bears in Minnesota: Implications for monitoring and management*, PloS ONE **5(8)** (2010) e12114.

[5] C. Guiver, D. Hodgson and S. Townley, *Positive state controllability of positive linear systems*, Systems and Control Letters **65** (2014) 23–29.

[6] T. B. Mayaka, J. D. Stigter, I. M. A. Heitknig and H. H. T. Prins, *A population dynamics model for the management of Buffon's kob (Kobus kob kob) in the Bénoué National Park Complex, Cameroon*, Ecological Modelling **176** (2004) 135–153.

[7] R. E. Miller and P. D. Blair, *Input-output analysis*, Cambridge University Press, Cambridge, 2009.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Strategies for colour mathematical morphology

Ángel Riesgo[1], Pedro Alonso[1], Irene Díaz[2] and Susana Montes[3]

[1] *Department of Mathematics, University of Oviedo, Spain*

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

[3] *Department of Statistics and O.R., University of Oviedo, Spain*

emails: ariesgo@yahoo.com, palonso@uniovi.es, sirene@uniovi.es, montes@uniovi.es

Abstract

Extending Mathematical Morphology from its original formulation for binary images to the case of colour images is not straightforward and there is still no standard approach. This paper discusses some criteria that can act as useful guidelines to classify the various existing techniques depending on how they deviate from a simpler reference greyscale model. In the final section, an approach with weight functions is presented.

1 Introduction

Mathematical Morphology (MM) is a robust theory used in image analysis [2, 5]. While there is now a standard theoretical framework for binary and greyscale images, its extension to colour remains problematic [2].

In the digital representation of images, colour values typically lie in a three-dimensional space adhering to a particular model like RGB (Red Green Blue) or HSL (Hue Saturation Luminance). We will mostly base our reasoning on HSL-style colour models, where the three dimensions match the distinct colour contributions of luminance, saturation and hue.

The goal of this article is twofold. We provide some definitions that can contribute to the classification of the existing approaches and we also propose some criteria for the parametrization of approaches that depend on a weight function.

This paper is organized as follows: Section 2 is dedicated to some basic definitions. Next, we assume that an ordering scheme that is closely correlated to luminance will yield better results and we establish two conditions for this. In Section 4 we introduce the most

  ISBN: 978-84-608-6082-2

basic colour ordering techniques and compare them against the conditions. Then we discuss a well-known modified form of lexicographical ordering. Finally, in Section 6 the ordering schemes that rely on weighting functions are discussed.

2 Basic definitions

We start by defining an image [7]:

Definition 2.1. An *image* is a map $A : \mathcal{D}_A \to \mathcal{C}$, where $\mathcal{D}_A \subset \mathbb{Z}^2$ and $\mathcal{C}$ is a bounded set. The subset $\mathcal{D}_A$ is the *definition domain* of the image and its elements are its *pixels*. If $\mathcal{C} = \{0, 1\}$, it is called a *binary image*; if $\mathcal{C}$ is a bounded subset of $\mathbb{R}$ (or $\mathbb{Z}$) such as $[0, 1]$ it is called a *greyscale image*. Finally, if $\mathcal{C}$ is a bounded subset of $\mathbb{R}^n$ (or $\mathbb{Z}^n$), with $n \geq 2$ it is called a *colour* or *multichannel image*.

The methodology of MM involves defining operators that transform an image. All such operators can be expressed in terms of two basic ones that are based on the idea of a moving probe that performs an operation on the pixel values. For the definition of these operations, we assume that the pixel values form a lattice [6], so any subset has a supremum and an infimum.

Definition 2.2. Let A be an image and B a binary image on a common definition domain $\mathcal{D}$. Let the *translation* T_z of an image A *by* $z \in \mathbb{Z}^2$ be a function that maps A to an image $T_z(A)(x) = A(x - z)$ if $x - z \in \mathcal{D}_A$ and 0 otherwise. The *dilation of* A *by* B is the function $\delta_B : A \to A$ defined by $\delta_B(A)(x) := \sup_{z \in \mathcal{D}: T_x(B)(z)=1}\{A(z)\}$. The binary image B is called the *structuring element (SE)* for the dilation.

Definition 2.3. Under the same conditions as in (2.2), the *erosion of* A *by the structuring element* B is the function $\epsilon_B : A \to A$ defined by $\epsilon_B(A)(x) := \inf_{z \in \mathcal{D}: T_x(B)(z)=1}\{A(z)\}$.

Figure 1 shows a binary image and its corresponding dilation and erosion images for a structural element consisting of a 3x3 square centred at the origin $(0, 0)$.

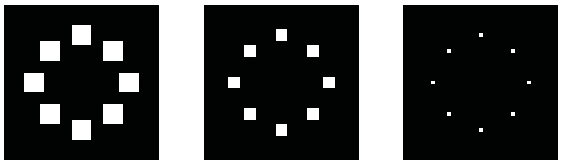


Figure 1: A binary image (centre) and its corresponding dilation (left) and erosion (right).

In the binary images, dilation extends the white artefacts (the *foreground pixels*) at the cost of the black background (the *background pixels*). This idea is readily generalised to the greyscale images, but the extension to colour is hampered by the multiple dimensions.

When referring to colour in an HSL-style space, we will use the term *luminance* for the L component. The other two components will jointly be referred to as the *chromaticity*. The one-dimensional subspace where the chromaticity cancels is the *line of greys*.

3 The perceptual role of luminance

In order to extend MM to colour in a way that preserves the behaviour on greyscale images, we will introduce some definitions that can help to classify the various techniques.

We shall assume that luminance plays a more significant role in the human perception of colour than chromaticity [3, p. 112]. As an illustration of this, figure 2 shows a decomposition of a photograph into its luminance channel and its chromatic part.

Figure 2: A photograph (left; by the authors) decomposed into its luminance channel in greyscale (centre) and its chromatic part at constant mean luminance (right).

Let us now formalise some definitions. A first obvious requirement is that colour MM should reduce to greyscale MM as a limit case.

Definition 3.1. Let $\mathcal{C} \subset \mathbb{R}^3$ be a colour space and let $\mathcal{G} \subset \mathcal{C}$ be its line of greys. A lattice structure defined on $\mathcal{C}$ is said to generate a *weakly greyscale-compatible* colour MM if the order it induces on $\mathcal{G}$ is the same as the total order on the corresponding greyscale space.

This is a first criterion for colour MM which allows us to discard any fanciful orderings that flout this condition. We call it *weak* as most of the common schemes satisfy it.

A second stronger requirement is based on making the MM operations commute with a projection of the colour values to a one-dimensional colour space.

Definition 3.2. Let $\mathcal{I}^c_\mathcal{D}$ be the set of colour images over a definition domain $\mathcal{D}$ with pixel values $\mathcal{C} \subset \mathbb{R}^3$ and let $\mathcal{I}^g_\mathcal{D}$ be the set of greyscale images over $\mathcal{D}$ with pixel values $\mathcal{G} \subset \mathbb{R}$. A *greyscale conversion* is a map $g : \mathcal{I}^c_\mathcal{D} \to \mathcal{I}^g_\mathcal{D}$.

Definition 3.3. Let $\mathcal{I}^c_\mathcal{D}$ be the set of colour images over $\mathcal{D}$ with pixel values in $\mathcal{C} \subset \mathbb{R}^3$ and let $g : \mathcal{I}^c_\mathcal{D} \to \mathcal{I}^g_\mathcal{D}$ be a greyscale conversion. A complete lattice structure defined on $\mathcal{C}$ is said to generate a *strongly greyscale-compatible colour MM* for g if dilations and erosions commute with g.

 ISBN: 978-84-608-6082-2

While this is too extreme a requirement in general, it is reasonable to expect that a powerful colour MM should not deviate too much from this idea.

4 The marginal and lexicographical ordering

A simple way of sorting multidimensional values is comparing one dimension and disregarding the others. V. Barnett [4] calls this *marginal ordering* or *M-ordering*. When working on a three-dimensional colour space XYZ, we can define three versions of M-ordering. Let p_1 and p_2 be two pixel values with coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2). Then we have:

Definition 4.1. *X-coordinate M-ordering:*

$$p_1 \leq_X p_2 \quad \text{if} \quad x_1 \leq x_2$$

The M-ordering is actually a preorder, which is not antisymmetric, so the choice of the supremum or infimum may lead to more than one possibility. This formal weakness can be fixed by comparing each coordinate in succession. This is the *lexicographical order*. In Barnett's classification, it is a form of *conditional ordering* (*C-ordering*) [4]:

Definition 4.2. *X-Y-Z lexicographical ordering:*

$$p_1 \leq_{lex_{XYZ}} p_2 \quad \text{if} \quad \begin{cases} x_1 \leq x_2 \\ \text{Or} \\ x_1 = x_2,\ y_1 \leq y_2 \\ \text{Or} \\ x_1 = x_2,\ y_1 = y_2,\ z_1 \leq z_2 \end{cases}$$

With an RGB colour model, the three possible M-orderings are weakly greyscale-compatible whereas in an HSL colour model only the L-coordinate M-ordering is. This leads us to discard M-orderings based on any of the chromaticity components.

The luminance M-ordering may lead to reasonable results for many types of colour images, like the simple ones in figure 3.

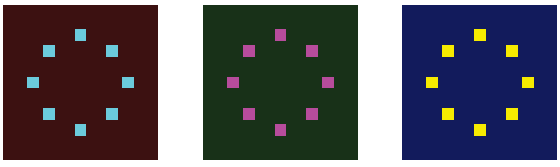


Figure 3: Three colour images where a luminance-based ordering would give the expected dilations and erosions.

 ISBN: 978-84-608-6082-2

For the lexicographical orderings, the six possibilities in the RGB colour model are all weakly greyscale-compatible. In the HSL model, the six possibilities are weakly greyscale-compatible too, but with an important caveat: if the luminance is relegated to the second or third place in the lexicographical cascade, the weak greyscale compatibility is unstable in the sense that a tiny modification in the chromatic part may completely change the ordering. This indicates that lexicographical orderings in the HSL colour space must prioritise the L coordinate for good behaviour.

The reasoning so far leads us to adopt a clear stance: of all the flawed marginal orderings, the luminance-based one, $\leq_L$, is best; of all the lexicographical orderings, the ones that prioritise luminance, $\leq_{lex_{Lxx}}$ are best. The strong condition of greyscale compatibility is only met by these marginal or conditional orderings that use the luminance channel as the first criterion. This reinforces our preference for such orderings.

5 Modified lexicographical ordering

A problem with the lexicographical ordering, however, is that it is affected too drastically by the variation of a single dimension. We can see this in figure 4. This is a problem with marginal or lexicographical orderings: while differences in a dimension, like luminance, may be more important than those affecting other dimensions, when the difference is too slight the other dimensions become more relevant.

The strong greyscale compatibility is a bad guideline for these cases and we must explore solutions that deviate from it. A compromise must be found where the ordering of the chromaticity part takes precedence when the luminance values are not too dissimilar.

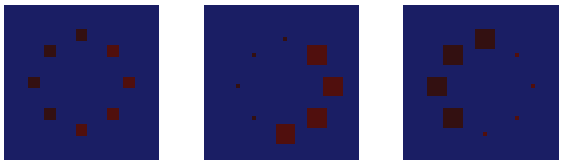


Figure 4: An image (left) and its corresponding dilation (centre) and erosion (right) based on an M-ordering by luminance. The tiny differences in luminance between the two groups of maroon squares lead to unexpected results.

Several variations of the lexicographical order that attempt to tackle this problem by quantising a colour dimension in discrete intervals have been proposed. E. Aptoula and S. Lefèvre give an excellent overview in [3]. A very successful scheme is the α-modulus lexicographical order due to J. Angulo and J. Serra [1], where the luminance values l are replaced by the integer part of l/α for the comparison. Using the ceiling function $\lceil x \rceil$ for the integer part, and assuming an HSL-style colour space where two pixels p_1 and p_2 have

 ISBN: 978-84-608-6082-2

coordinates (l_1, s_1, h_1) and (l_2, s_2, h_2), we have:

Definition 5.1. *L_α-S-H lexicographical ordering:*

$$p_1 \leq_{L_\alpha} p_2 \quad \text{if} \quad \begin{cases} \left\lceil \frac{l_1}{\alpha} \right\rceil \leq \left\lceil \frac{l_2}{\alpha} \right\rceil \\ \text{Or} \\ \left\lceil \frac{l_1}{\alpha} \right\rceil = \left\lceil \frac{l_2}{\alpha} \right\rceil, \; s_1 \leq s_2 \\ \text{Or} \\ \left\lceil \frac{l_1}{\alpha} \right\rceil = \left\lceil \frac{l_2}{\alpha} \right\rceil, \; s_1 = s_2, \; h_1 \leq h_2 \end{cases}$$

The effect of the α parameter is to smooth out the luminance comparisons so that the possible values are grouped into equivalence categories when they lie within the same luminance interval. This can similarly be extended to the second dimension as an L_α-S_β-H lexicographical ordering.

In the α-modulus lexicographical ordering, α is a constant although the effect of the chromaticity is more marked for luminance values at the middle of the range than near the black or white ends. An improvement would be to make α vary across the luminance values. This was pointed out by E. Aptoula and S. Lefèvre [3], who provided their own variant of this ordering where α is a function:

Definition 5.2. *$L_{\alpha(L)}$-S-H lexicographical ordering:*

$$p_1 \leq_{L_{\alpha(L)}} p_2 \quad \text{if} \quad \begin{cases} \left\lceil \frac{l_1}{\alpha(L)} \right\rceil \leq \left\lceil \frac{l_2}{\alpha(L)} \right\rceil \\ \text{Or} \\ \left\lceil \frac{l_1}{\alpha(L)} \right\rceil = \left\lceil \frac{l_2}{\alpha(L)} \right\rceil, \; s_1 \leq s_2 \\ \text{Or} \\ \left\lceil \frac{l_1}{\alpha(L)} \right\rceil = \left\lceil \frac{l_2}{\alpha(L)} \right\rceil, \; s_1 = s_2, \; h_1 \leq h_2 \end{cases}$$

But a bad side-effect of the tolerance in the luminance comparison introduced by both types of α-modulus lexicographical orderings is that the weak greyscale compatibility is lost. The next section explores an alternative approach.

6 Colour ordering by a weighting function

The colour space can also be given a lattice structure through the association of each colour vector with a one-dimensional weight value. We first give a name to such functions:

Definition 6.1. Given a colour space $\mathcal{C} \subset \mathbb{R}^3$, a function $w : \mathcal{C} \to W$, where W is a bounded subset of $\mathbb{R}$ (or $\mathbb{Z}$ in a digital representation), is called a *colour-weighting function*.

The choice of a weighting function results in a preorder in the colour space. For better mathematical behaviour, this preorder can be turned into a total order through the use of a lexicographical cascade as a second criterion when two colour values have the same weight.

Definition 6.2. *w-weighted ordering:*

$$p_1 \leq_{w_\alpha} p_2 \quad \text{if} \quad \begin{cases} w(p_1) \leq w(p_2) \\ \text{Or} \\ w(p_1) = w(p_2),\ p_1 \leq_L p_2 \end{cases}$$

Most often we will be interested in linear functions with normalised coefficients:

Definition 6.3. Given a colour space $\mathcal{C} \subset \mathbb{R}^3$, a colour-weighting function $w : \mathcal{C} \to W$ is called a *normalised linear weighting function* if for any pixel value $p = (l, s, h)$, $w(p) = w_L l + w_S s + w_H h$ with $w_L, w_S, w_H \in [0, 1]$ and $w_L + w_S + w_H = 1$.

The lexicographical ordering can be approximated by a normalised linear weighting function that gives significantly different weights to each component. For example, if our colour space has three coordinates l, s and h, all of them in the $[0, 1]$ real interval, sorting the colour values with a weighting function such as $w = 0.99l + 0.0099s + 0.0001h$ will essentially give the same results as the *L-S-H* lexicographical ordering.

Similarly, the threshold parameter α can be adapted to the language of weighting functions. Assuming that the ranges of the l, s and h coordinates are in the $[0, 1]$ real interval, if we want $\alpha \in (0, 1)$ to be the threshold such that a pixel value $p_1 = (l_1, s_1, h_1)$ will always be smaller than $p_2 = (l_1 + \alpha, s_2, h_2)$, then we have that $w_L l_1 + w_S s_1 + w_H h_1 \leq w_L(l_1 + \alpha) + w_S s_2 + w_H h_2$ for any values of s_1, s_2, h_1 and h_2, which leads to $w_S + w_H \leq w_L \alpha$ and, consequently $w_L \geq \frac{1}{1+\alpha}$. So, a weight of $\frac{1}{1+\alpha}$ plays a similar threshold role as α in the α-modulus lexicographical orders. But the situation is not the same, as the weighted approach will still account for small differences in the first coordinate and preserve the weak greyscale compatibility.

An improvement here consists in making the weights vary with the luminance coordinate, in a similar way to what E. Aptoula and S. Lefèvre proposed [3] for the α-modulus lexicographical order. Aptoula and Lefèvre's algorithm calculates the luminance equivalence classes starting from a priority distribution $f : \mathbb{N} \to [0, 1]$ that determines how the $[0, 1]$ is to be stretched, with the identity distribution $f(x) = x$ giving Angulo and Serra's α-modulus lexicographical ordering.

We can easily build a similar algorithm to make the weights depend on the luminance of the pixel value. For implementation purposes, we should assume that the luminance coordinate is digitised in N integer values $\{0, \ldots, N - 1\}$. If we assume that a priority distribution $f : \mathbb{N} \to [0, 1]$ determines the relative sensitivity to changes along the $[0, 1]$ range, then algorithm (1) builds an array of weight values for each luminance value.

Algorithm 1 Function that builds an array of luminance weights

```
function BuildLuminanceWeightArray(alpha, f)
    luminanceWeightArray ← [0, ^{N−1}⋯, 0]
    for all i ∈ [0, ..., N − 2] do
        variableAlpha ← alpha ∗ (f(i + 1) − f(i))
        luminanceWeight ← 1/(1 + variableAlpha)
        luminanceWeightArray[i] ← luminanceWeight
    end for
    return luminanceWeightArray
end function
```

Once this array of weights has been initialised with (1), the colour weight for a pixel value is calculated by first checking the luminance value l as an N-bit integer number and then accessing the l-th element in the array as the weight w_L. This mechanism can similarly be extended to the other colour dimensions.

Acknowledgements

The research in this paper has been supported in part by grant MINECO-TIC2014- 59543-P; its financial support is gratefully acknowledged.

References

[1] J. Angulo, J. Serra, *Morphological Coding of Color Images by Vector Connected Filters*, In IEEE Proceedings of the 7th International Symposium on Signal Processing and its Applications, 2003, Vol. 1, pages 69–72.

[2] E. Aptoula, S. Lefèvre, *A comparative study on multivariate mathematical morphology*, Pattern Recognition **40 (10)** (2007) 2914–2929.

[3] E. Aptoula, S. Lefèvre, *On lexicographical ordering in multivariate mathematical morphology*, Pattern Recognition **29 (2)** (2008) 109–118.

[4] V. Barnett, *The ordering of multivariate data*, Journal of the Statistical Society. Series A (general) **139 (3)** (1976) 318–355.

[5] C. Ronse, L. Najman, E. Decencière (Eds.), *Mathematical Morphology: 40 Years On: Proceedings of the 7th International Symposium on Mathematical Morphology*, Springer Verlag, Dordrecht, the Netherlands, 2005.

[6] C. Ronse, *Why mathematical morphology needs complete lattices*, Signal Processing **21 (2)** (1990) 129–154.

[7] P. Soille, *Morphological Image Analysis: Principles and Applications*, Springer-Verlag, Berlin Heidelberg, 2004.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Assistance Management Application based on IBSC for Emergency Situations

Alexandra Rivero-García[1], Candelaria Hernández-Goya[1], Iván Santos-González[1] and Pino Caballero-Gil[1]

[1] *Department of Computer Engineering and Systems, Universidad de La Laguna, Tenerife, Spain.*

emails: ariverog@ull.edu.es, mchgoya@ull.edu.es, jsantosg@ull.edu.es, pcaballe@ull.edu.es

Abstract

A system designed to serve the greatest number of injuries using the shortest possible time and cost in extreme situations is described. It is composed by a mobile application and a web service. The mobile application is devoted to medical staff providing them with the location of victims on a map as well as an assistant indicating the route to follow to care for them based on the severity of their condition. Doctors may also use a functionality of the application to contact peers through a video call when additional help is needed.

The proposal combines an HMAC scheme to protect NFC tags and an Identity-Based Signcryption scheme for communication confidentiality, authenticity and integrity both among peers and between server and medical staff.

Key words: IBSC, HMAC, Security, Triage, Emergency

1 Introduction

The fast evolution of communication technologies and smartphones can help in many complex scenarios. These devices are used to support different daily tasks thanks to their small size and high performance.

The starting point of this project was the application developed in [1], where a mobile system for victims classification in emergency situations was implemented. This paper presents a distributed platform for improving logistics of medical staff in emergency situations based on data obtained from the triage application aforementioned.

A commonly accepted triage definition follows. A simple, completed, objective and fast process to obtain an initial clinical assessment of people with the objective of evaluating their immediate survival capacities and prioritizing them according their severity is a triage. In order to achieve the classification, all triage systems distinguish two steps. The first triage or simple triage is used for the generation of a classification based on the severity of injuries of the victims evaluating their survival skills in some seconds. The second triage is where medical staff analyze each patient's state: bruises, wounds and injuries.

In this work START (Simple Triage and Rapid Treatment Algorithm) method is used as first triage. Its output is the victim's classification based on coloured tags, where each colour defines the priority of the victim: black, dead or irrecoverable victims; red, victims requiring immediate care; yellow, victims requiring urgent care but who can wait for treatment from half an hour to one hour; green, victims who are not seriously injured. They can wait for treatment more than an hour. Here the use of NFC communication is proposed to deal with the triage result. NFC (Near Field Communication) stickers are used to save triage results based on the generation of a HMAC scheme. Furthermore, the route to attend victims for each doctor is shown through a map in their smartphones based on the priorities of victims and they can share information P2P (peer-to-peer) with their colleagues in the affected area. All these communications are protected through an IBSC scheme.

2 Preliminaries

The integration of new technologies into emergency situations management and medical care has allowed the development of tools that help to the coordination between medical staff in emergency scenarios. There are different proposals designed to help to find missing persons after a large-scale disaster. Such as People Locator and ReUnite [2], Google person finder [3] and Safety Check of Facebook [4]. All these systems try to verify and share the status of people after some disaster, specifically the proposal of Facebook share all the information with the victim's friends in this social network.

There are some applications for smartphones for emergency response. "Fire Department" [5] is one of the most well known application from the San Ramon Valley Fire Protection District in USA. This application generate a map with the recent emergency events and in the same map users can find the location of an automated defibrillator. CodeBlue system [6], is a U.S. Army system, centered around the use of wireless sensors in emergencies. This is a combined hardware and software platform for medical wireless sensor networks that provides wireless monitoring and tracking of patients and first responders [7]. These application are not designed for large-scale emergency scenarios.

The SAHANA[8] foundation project aims to provide a set of modular, web-based disaster management applications. This project includes tools for synchronization between multiple instances: a Missing Person Registry, Request and Pledge Management System

and Volunteer coordination. Because this proposal is a web-based framework, it has the problem of relying on communication to the centralized web-server, and thus can not take advantage of mobile nodes.

The unique identification of victims is a requirement for any emergency triage. Barcodes are usually used thus facilitating mechanical reading [9]. Barcodes are cheap and easy to create, they can be generated just using a standard printer. But in an emergency situation having a printer in the affected zone is not realistic. RFID (Radio Frequency IDentification) is a very useful technology for victim identification as it is explained in [10] and in [11]. Two types of tags exists, passive tags, that use the energy received from the reader to send the identifier, and active tags, that include a battery to increase its distance range. The problem of this kind of communication is that a RFID reader is needed and no security tools were provided. In [12] a specific triage tag technology is proposed. These electronic triage tags use noninvasive biomedical sensors to continuously monitor the vital signs of a patient and deliver pertinent information to first responders. These are not triage tag for emergency situations.

3 Global View

The overall objective of the proposal is to generate a tool to save time in emergency situations. Thus doctors have a map that help them in every moment to decide the route to patients based on the severity of the injuries. In this way, collisions of doctors to assist the same patient are avoided and decisions are taken based on priority.

There are two stages in the route generation. The first one consists on the revision of the affected area. The first respondents in the disaster apply START triage method obtaining a victims' classification based on coloured tags. In this case NFC [13] tags, specifically NFC stickers are used to store the triage result. Each triage has a location in the central server.

The second stage is based on the victim's attention taking into account the results of the triage priorities. In this step the victims' locations given by the first triage is essential being the starting point. A graph of each colour is generated based on the victim's location in a map. Nodes represent victims and edges are the routes to reach them. Each edge has a cost. This cost is the distance between two nodes. In this way, if we have a situation like figure 1a, the resulted graph is figure 1b.

All information related to the patients who must be attended by a doctor is done through a mobile application. It indicates to the medical staff through a map his/her current location and the next patient to assist.

The application has enabled a feature called "emergency support". With this function when a doctor or nurse requires additional help from peers he/she can activate this mode. When they activate this feature all health personnel in the affected area receives the notification and simply by clicking on it, they can start a video call to help his/her

(a) First triage solution. (b) Graph of each triage colour. (c) Triage route for each doctor.

colleague.

This functionality is designed to support healthcare workers and improve the use of time in transfers between patients. Note that this feature opens a communication channel between two partners through a video streaming. Due to the high amount of information exchanged the connection will take place by LTE (Long Term Evolution)[14], specifically LTE-Direct to ensure adequate and secure communication between nodes that connect.

Whenever a doctor has just treated a victim, he/she can take his/her mobile, read the tag, mark the point as completed and check next victim status. When the doctor arrives to the location of the new victim, the node is automatically marked on the map as being in the care process but he/she can read the sticker to be sure of the authenticy of the node. The period devoted to reach a new node is called "travelling time". Doctors can receive notifications called "emergency support" when they are in this "travelling time" to avoid constant notifications that may mislead the staff in the middle of an assistance.

4 Decision-making system

In the generation of doctors' routes an undirected graph is created from the points defined during the triage. There are as many points as patients on the map, these are the vertices of our graph. The edges will be defined according to the closest path between the vertices. This distance between points will be the cost of the edge.

Once the graph is generated, the amount of resources and the place where they are are needed. In this case resources are doctors (number of doctors $\#d$) that will assist patients. Their position at all time is known (figure 1c).

The initial step is assigning to each doctor the closest red node. This is based on the Greedy Algorithm [15] taking into account different priorities. Several situations can be found. Note if there are red nodes (number of red nodes $\#r$) the other colours are not considered. When these victims are attended the yellow nodes(y) are taken into account and finally green nodes(g).

When $\#d > \#r$ each patient with red tag is assigned to a doctor, and then they continue with the next highest priority node. That is, once the doctors are assigned to all patients with red tag classification, every "free" doctor is assigned to the closest yellow node.

If $\#d \leq \#r$ $Sum(m)Sum(r)$: When we have less doctors than priority nodes routes based on priorities are generated, that is, a graph is generated for each member of the medical staff based on red nodes and forgetting the other nodes. When red nodes are finished, the system generates the graph with yellow nodes and, in the end, with greens.

Hence a graph of routes is generated for each doctor based on the combination of different sub-graphs that are generated with patients of the same priority level. At the time of generating the graph of the following categories (colours) the last vertex added to the previous graph is the starting point of the new graph.

This generation separated sub-graphs is because by regulations when applying triage schemes patients may be attended in this order of their severity.

The incorporation of new medical staff or new casualties does not cause any problems or additional cost. If more nodes are added to the graph the doctors' routes are updated paying attention to the new characteristics of the affected area. The routes will be reseted and each doctor can continue his/her work without worrying about such distractions. Given a constraint, doctors who are in the "travelling time" will not receive the route update until he/she has attend next victim, this will the starting point of the route.

5 Victims identification: NFC tags and HMAC scheme

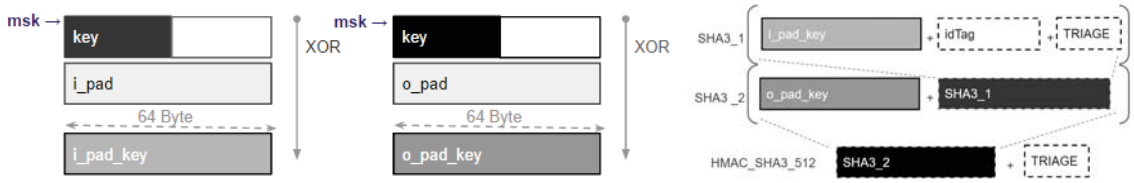
In the system proposed, a member of the medical staff assigns NFC tags to victims. These tags contain the result of the triage, that is to say the color of the triage classification, jointly with the result of a HMAC generated by the server, the physical identifier of the NFC tag (idTag) and some server data explained later in the paper. The stored information will serve as patient identification both in for triage as well as in the medical records generated later on at the hospital.

The identification of patients through NFC stickers is realized by using the mobile phone as an NFC reader. The smartphone sends the physical tag identifier to the server and there two 64 bytes arrays are generated as in [16]. They are *ipad* and *opad* arrays, and they have default values defined at the initialization stage. The new arrays are generated through a XOR operation combining the previous values and the Master Private Key (mpk). The results are *ipadk* y *opadk* arrays (figure 2a). After that, the HMAC value is generated with the physical tag identifier and the triage colour result T_{result} (it is a letter for each colour: B, black; R, red; Y, yellow and G, green), so the system applies a hash function to the concatenation of these fields and the *ipadk*. The output of this hash concatenated with the *opadk* is the input to another hash function.

The global function may be described as:

$$HMAC(Tid, mpk) = HASH((opadk)||HASH(ipadk)||idTag||T_{result})$$

The hash function chosen for the implementation is a $SHA3_{512}$. The final output will be the identifier that the will be saved in the NFC sticker tag, as you can see in figure 2b.



(a) HMAC Keys Generation

(b) HMAC Hash Operation

In the affected zone, when a doctor or a member of the medical staff want to access to the triage result of a patient, he/she has to read the NFC sticker through the mobile application which sends the data of the physical tag identifier and HMAC to the server. The server is who verifies the authenticity of the tag and generates a new node. The doctors can see all the nodes in the mobile phone, specifically the nodes on his/her routes.

6 Medical Staff security: IBSC scheme

Two communication modes are supported related with medical staff, the communication with the server (to check NFC tags authenticity and synchronyzing routes) and the communication peer-to-peer (video-calls and chats). Authentication against the server and peers and integrity of shared data is included.

In order to achieve an Identity-Based Signcryption scheme (IBSC) is used. This complex cryptosystem is a combination of IBE (Identity-Based Encryption) and IBS (Identity-Based Signature) that provides private and authenticated delivery of information between two parties in an efficient way with a composition of an encryption scheme with a signature scheme [17]. This approach offers the advantage of simplifying management by not having to define a public key infrastructure. This type of scheme was chosen due to its low computational complexity, efficiency in terms of memory and its usability. The communication flow is shown in figure 3.

A crucial part of the proposal is a Private Key Generator (PKG), a server in charge of generating health staff private keys. The identifier of medical staff is the number of registered medical practitioners and for nurses the same(ID). Next, we describe the mathematical basic tools used as well as the notation included in their description.

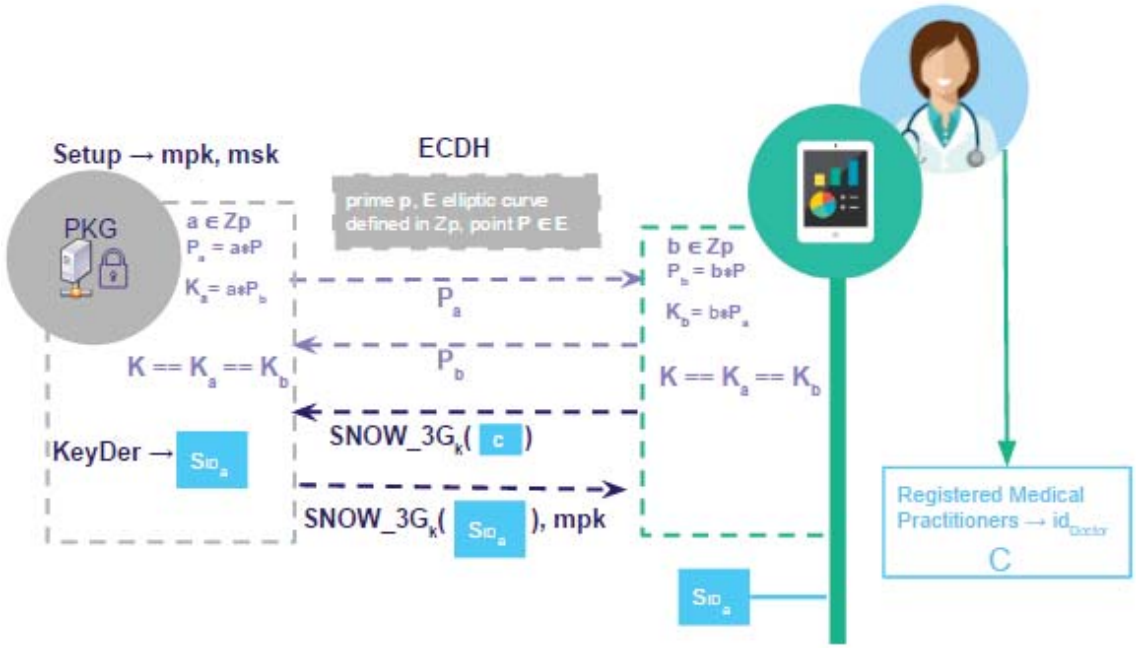


Figure 3: Private Key Exchange

Definition 1. *Considering two cycling groups* $(G, +)$ *and* $(V, \cdot)$ *of the same prime order* q. P *is a generator of* G *and there is a bilinear map paring* $\hat{e} : G \times G \to V$ *satisfying the following conditions:*

- *Bilinear:* $\forall P, Q \in G$ *and* $\forall a, b \in \mathbb{Z}$, $\hat{e}(aP, bQ) = \hat{e}(P, Q)^{ab}$
- *Non-degenerate:* $\exists P_1, P_2 \in G$ *that* $\hat{e}(P_1, P_2) \neq 1$. *This means if* P *is generator of* G, *then* $\hat{e}(P, P)$ *is a generator of* Q.
- *Computability: there exists an algorithm to compute* $\hat{e}(P, Q), \forall P, Q \in G$

Some hash functions denoted as follows are also needed:

$$H_1 : \{0,1\}^* \to G^*, H_2 : \{0,1\}^* \to \mathbb{Z}_q^*, H_3 : \mathbb{Z}_q^* \to \{0,1\}^n$$

where the size of the message is defined by n. The signcryption scheme used is the IDSC (Identity-Based Signcryption Scheme) proposed in [18]. Next we describe some basic notation used: $x \xleftarrow{r} S$ stands for an element x randomly selected from a set S, $x \leftarrow y$ denotes the assignation of the value y to x and $||$ is used for concatenation.

The steps needed for the signcryption scheme are the following:

- **SETUP:** The initial parameters are established and the server generates the master public key (mpk, represented as Q_{TA}) and the master secret key (msk, represented as t). For that a prime q based on some private data $k \in \mathbb{Z}$, two groups G and V of order q and a bilinear pairing map $\hat{e} : G \times G \to V$ are selected. $P \in G$ is selected randomly and the hash functions H_1, H_2 and H_3 are also chosen.
- **EXTRACT (ID):** In this step the secret key for each member of the medical staff based on their ID is generated. The public key $Q_{ID} \in G$ and the secret key $S_{ID} \in G$ are calculated taking into account the master private key (t). It should be pointed out that this key exchange between server and the doctor is performed using the stream cipher SNOW3G [19] under the session key obtained through an ECDH (Elliptic Curve Diffie-Hellman)[20]. This data exchanged is critical to the safety of following connections as you can see in figure 3.
- **SIGNCRYPTION (S_{ID_a}, ID_b, m):** All the messages $m \in \{0,1\}^n$ will be encrypted and signed. The receiver's public key is generated taking into account ID_b and then the message is signed with S_{ID_a} and encrypted with Q_{ID_b} giving as result σ (a t-uple of three components: c, T, U).
- **UNSIGNCRYPTION (ID_a, S_{ID_b}, σ):** If everything is right, the message $m \in \{0,1\}^n$ is returned. Otherwise, if there are some problems in the signature or in the encryption of m, $\perp$ is returned. The sender's public key is generated taking into account ID_a and then the message is unencrypted with S_{ID_b}.

A formal description of all these steps follows.

SETUP

$$t \xleftarrow{r} \mathbb{Z}_q^*$$

$$Q_{TA} \leftarrow tP$$

EXTRACT

$$Q_{ID} \leftarrow H_1(ID)$$

$$S_{ID} \leftarrow tQ_{ID}$$

SIGNCRYPTION

$$Q_{ID_b} \leftarrow H_1(ID_b)$$
$$x \xleftarrow{r} \mathbb{Z}_q^*$$
$$T \leftarrow xP$$
$$r \leftarrow H_2(T||m)$$
$$W \leftarrow xQ_{TA}$$
$$U \leftarrow rS_{ID_a} + W$$
$$y \leftarrow \hat{e}(W, Q_{ID_b})$$
$$k \leftarrow H_3(y)$$
$$c \leftarrow k \oplus m$$
$$\sigma \leftarrow (c, T, U)$$

UNSIGNCRYPTION

$$Q_{ID_a} \leftarrow H_1(ID_a)$$
$$split \quad \sigma \quad as \quad (c, T, U)$$
$$y \leftarrow \hat{e}(S_{ID_b}, T)$$
$$k \leftarrow y$$
$$m \leftarrow k \oplus c$$
$$r \leftarrow H_2(T||m)$$

Verification:

$$\hat{e}(U, P) == \hat{e}(Q_{ID_a}, Q_{TA})^r \cdot \hat{e}(T, Q_{TA})$$

Note: if the verification is successfull m is returned, otherwise $\perp$ is returned.

7 Security Analysis

The proposed scheme provides protection against different attacks. Here we describe some of them.

A spoofing attack and/or cloning of the card will be hardly successful since it would involve the generation of the HMAC described taking into account the master key of the server and the ID card. Even if an outsider obtains this information, it should be noted that the physical identifier of a NFC tag is unique to each element.

If someone want to emulate the card from an Android device, the emulated device goes from being passive to being active. So the attack would be detected since the application has the restriction that only read NFC tags that are passive. At the time of its implementation in Android are different and completely distinguishable communications.

The Denial of Service (DoS) attacks relating to make requests to the server are restricted because only requests associated with a number of legitimate members of the medical staff will take effect. Once the corresponding private key is assigned, more requests of this kind will be not attended.

A "Man in the Middle" attack conveys a successful authentication to the server with an identifier of legitimate members of the medical staff. This false identification would be easily detectable because the number of members who can make requests to the server is limited to those who are working at the time of the request.

8 Conclusions and Future Work

A system that may improve logistics and attention of casualties in extreme situations has been presented. The priority is to serve the greatest number of injuries using the shortest possible time and cost. The tool consists on a mobile application and a web service. The mobile application helps health staff to know in every moment the position of the victim and where they must go. Doctors may us the "emergency support" tool to contact peers through a video call when they require additional support. Data security is a key objective for this reason a HMAC scheme is used to protect NFC tags and an Identity-Based Signcryption for the communications. A first approach has been implemented in Android and Nodejs with NFC tags. More functionalities should be added to the server such as statistics, a real-time map with events, etc.

Acknowledgements

Research supported by TESIS2015010102, TESIS2015010106, RTC-2014-1648-8, TEC2014-54110-R, MTM-2015-69138-REDT and DIG02-INSITU.

References

[1] A. Rivero-Garcıa, C. Hernández-Goya, I. Santos-González, and P. Caballero-Gil, "Fast-triaje: A mobile system for victim classification in emergency situations."

[2] "People locator and reunite web page." [Online]. Available: https://lpf.nlm.nih.gov/PeopleLocator-ReUnite

[3] "Google person finder," a google.org project. [Online]. Available: https://lpf.nlm.nih.gov/PeopleLocator-ReUnite

[4] "Facebook safety check." [Online]. Available: https://www.facebook.com/about/safetycheck/

[5] "Fire protection district application." [Online]. Available: http://www.firedepartment.org

[6] K. Lorincz, D. J. Malan, T. R. Fulford-Jones, A. Nawoj, A. Clavel, V. Shnayder, G. Mainland, M. Welsh, and S. Moulton, "Sensor networks for emergency response: challenges and opportunities," *Pervasive Computing, IEEE*, vol. 3, no. 4, pp. 16–23, 2004.

[7] "Code blue platform." [Online]. Available: http://codeblue.com/applications/

[8] "Sahana software fundation." [Online]. Available: http://sahanafoundation.org/products/

[9] M. Neuenschwander, M. R. Cohen, A. J. Vaida, J. A. Patchett, J. Kelly, and B. Trohimovich, "Practical guide to bar coding for patient medication safety," *AMERICAN JOURNAL OF HEALTH SYSTEM PHARMACY*, vol. 60, no. 8, pp. 768–779, 2003.

[10] S. Inoue, A. Sonoda, K. Oka, and S. Fujisaki, "Emergency healthcare support: Rfid-based massive injured people management," in *Proceedings of the fourth International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications, Irvine, CA*, 2006.

[11] "Baracoda. idbluean efficient way to add rfid reader/encoder to bluetooth pda and mobile phones." [Online]. Available: http://www.baracoda.com/baracoda/products/p21.html

[12] T. Gao, T. Massey, L. Selavo, D. Crawford, B.-r. Chen, K. Lorincz, V. Shnayder, L. Hauenstein, F. Dabiri, J. Jeng *et al.*, "The advanced health and disaster aid network: A light-weight wireless medical system for triage," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 1, no. 3, pp. 203–216, 2007.

[13] R. Want, "Near field communication," *IEEE Pervasive Computing*, no. 3, pp. 4–7, 2011.

[14] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley Online Library, 2009.

[15] B. M. Waxman, "Routing of multipoint connections," *Selected Areas in Communications, IEEE Journal on*, vol. 6, no. 9, pp. 1617–1622, 1988.

[16] M. Bellare, R. Canetti, and H. Krawczyk, "Message authentication using hash functions: The hmac construction," *RSA Laboratories CryptoBytes*, vol. 2, no. 1, pp. 12–15, 1996.

[17] X. Boyen, *Identity-based signcryption*. Springer, 2010.

[18] J. Malone-Lee, "Identity-based signcryption." *IACR Cryptology ePrint Archive*, vol. 2002, p. 98, 2002.

[19] I. Santos-González, A. Rivero-García, P. Caballero-Gil, and C. Hernández-Goya, "Alternative communication system for emergency situations." in *WEBIST (2)*, 2014, pp. 397–402.

[20] V. S. Miller, "Use of elliptic curves in cryptography," in *Advances in Cryptology-CRYPTO85 Proceedings*. Springer, 1985, pp. 417–426.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Stability and 0ptimal Control of a Delayed HIV Model

Diana Rocha[1], Cristiana J. Silva[1] and Delfim F. M. Torres[1]

[1] *Center for Research and Development in Mathematics and Applications (CIDMA)*
Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

emails: diana.isa.rocha@ua.pt, cjoaosilva@ua.pt, delfim@ua.pt

Abstract

We propose and investigate a delayed model that studies the relationship between HIV and the immune system during the natural course of infection and in the context of antiviral treatment regimes. Sufficient criteria for local asymptotic stability of the infected equilibrium and the viral free equilibrium are given. We propose and analyse an optimal control problem with time delay in the state variables, where the objective is to find the optimal treatment strategy that maximizes the number of $CD4^+$ T cells as well as the CTL (immune response cells), keeping the cost as low as possible.

*K*eywords: HIV modelling, time delay, stability, optimal control.

*M*SC 2010: 34C60, 49K15, 92D30.

1 Introduction

The study of mathematical models for human immunodeficiency virus (HIV) infection is a subject of current interest, both at population and cell level (see, e.g., [6, 9, 10] and references cited therein). Based on the model of [10], in this work we analyse a mathematical model that studies the relationship between HIV and the immune system during the natural course of infection and in the context of antiviral treatment regimes. The model considers four variables: uninfected $CD4^+$ T cells (x), infected $CD4^+$ T cells (y), cytotoxic T lymphocyte (CTL) precursors (CTLp) (w), and CTL effectors (z). Uninfected $CD4^+$ T cells are produced at a rate λ, die at a rate d, and become infected by free virus at a rate b. Infected cells decay at a rate a and are killed by CTL effectors at a rate p. Thus, proliferation of the CTLp population is given by $cxyw$ and is proportional to both virus load

 ISBN: 978-84-608-6082-2

(y) and the number of uninfected T helper cells (x). CTLp die at a rate b and differentiate into effectors at a rate cq. CTL effectors die at a rate h. Mathematically, the model under investigation is described by

$$\begin{cases} \dot{x} = \lambda - dx - \beta xy, \\ \dot{y} = \beta xy - ay - pyz, \\ \dot{w} = cyw(1-q) - bw, \\ \dot{z} = cqyw - hz. \end{cases} \tag{1}$$

Time delay plays an important role in the dynamics of HIV infection, see, e.g., [7, 8] and references therein. In this work, we introduce a discrete time-delay into model (1) that represents the incubation period. We prove local asymptotic stability of the viral free and infected equilibriums, for any time delay. The stability results are illustrated through some numerical simulations.

Optimal control theory has been applied to HIV models, see, e.g., [1, 3] and references therein. In our work, we propose and solve an optimal control problem where the objective is to find the optimal treatment strategy that maximizes the number of $CD4^+$ T cells as well as the CTL (immune response cells), keeping the cost, measured in terms of chemotherapy strength and a combination of duration and intensity, as low as possible.

2 Model with time delay

In epidemiological literature, a latent or incubation period is often modelled by incorporating it as a delay effect [4]. We consider the following delayed model, where τ ($\tau > 0$) represents the incubation period:

$$\begin{cases} \dot{x} = \lambda - dx(t) - \beta x(t)y(t), \\ \dot{y} = \beta x(t-\tau)y(t-\tau) - ay(t) - py(t)z(t), \\ \dot{w} = cy(t)w(t)(1-q) - bw(t), \\ \dot{z} = cqy(t)w(t) - hz(t). \end{cases} \tag{2}$$

The initial conditions for system (2) are

$$x(\theta) = \varphi_1(\theta),\ y(\theta) = \varphi_2(\theta),\ w(\theta) = \varphi_3(\theta),\ z(\theta) = \varphi_4(\theta), \tag{3}$$

$-\tau \leq \theta \leq 0$, where $\varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)^T \in C$ with C the Banach space $C\left([-\tau, 0], \mathbb{R}^4\right)$ of continuous functions mapping the interval $[-\tau, 0]$ into $\mathbb{R}^4$. The usual local existence, uniqueness and continuation results apply [2, 5]. Therefore, there exists a unique solution $(x(t), y(t), w(t), z(t))$ of (2) with initial conditions (3), for all time $t \geq 0$. From biological meaning, we further assume the initial functions to be non-negative:

$$\varphi(\theta) \geq 0, \quad \text{for} \quad \theta \in [-\tau, 0], \quad i = 1, \ldots, 4.$$

From [11, Theorem 2.1] all solutions of (2) satisfying (3) are bounded for all time $t \geq 0$, which ensures not only local existence, proved applying results from [2, 5], but also the existence of a solution for all time $t \geq 0$. The viral-free equilibrium is given by

$$E_0 = \left(\frac{\lambda}{d}, 0, 0, 0\right)$$

and the infected equilibrium is given by

$$E_+ = \left(\frac{-c\lambda(q-1)}{b\beta + cd(1-q)}, \frac{b}{c(1-q)}, \frac{h(q-1)(ab\beta + acd - \beta c\lambda + \beta c\lambda q - acdq)}{bpq(b\beta + cd(1-q))}, \right.$$
$$\left. \frac{\beta c\lambda(1-q)}{b\beta p + cdp(1-q)} - \frac{a}{p}\right).$$

3 Local stability of the delayed model

We prove the following result for the viral free equilibrium E_0.

Theorem 1 *If $da > \beta\lambda$, then the viral free equilibrium E_0 of (2) is locally asymptotically stable for any time delay $\tau \geq 0$.*

We also show the local stability of the endemic equilibrium E_+ for any time delay.

Theorem 2 *If $da < \beta\lambda$, then the infected equilibrium E_+ of (2) is locally asymptotically stable for any time delay $\tau \geq 0$.*

These analytical results are confirmed through numerical simulations.

4 Optimal control problem with state delays

We can introduce drug therapy into the model by assuming that treatment reduces the rate of viral replication, expressed as $s\beta xy$, where $0 < s < 1$. The drugs are 100% efficient if $s = 0$ and have no effect if $s = 1$ [10]. In our paper, we consider that $s = s(t)$, $t \in [0, t_f]$, is a control function that represents the efficacy of treatment. Our aim is to find the optimal treatment strategy $s(t)$ that maximizes the number of $CD4^+$ T cells x as well as the CTL z (immune response cells), keeping the cost low (measured in terms of chemotherapy strength and a combination of duration and intensity). We propose the following control system, with the discrete non-negative time delay τ in the state variables:

$$\begin{cases} \dot{x}(t) = \lambda - dx(t) - s(t)\beta x(t)y(t), \\ \dot{y}(t) = s(t)\beta x(t-\tau)y(t-\tau) - ay(t) - py(t)z(t), \\ \dot{w}(t) = cy(t)w(t)(1-q) - bw(t), \\ \dot{z}(t) = cqy(t)w(t) - hz(t). \end{cases} \tag{4}$$

 ISBN: 978-84-608-6082-2

The initial conditions for the state variables w and z and, due to the delays, initial functions for the state variables x and y, are given by:

$$\begin{aligned} &w(0) = w_0 \geq 0,\ z(0) = z_0 \geq 0, \\ &x(t) = x_0 \geq 0 \quad \text{for} \quad -\tau \leq t \leq 0, \qquad y(t) = y_0 \geq 0 \quad \text{for} \quad -\tau \leq t \leq 0. \end{aligned} \tag{5}$$

We consider the set of admissible control functions given by

$$\Theta = \left\{ s(\cdot) \in L^{\infty}(0, t_f) \,|\, 0 \leq s(t) \leq 1\,,\, \forall\, t \in [0, t_f] \right\} \tag{6}$$

and the L^1 objective functional

$$J(s(\cdot)) = \int_0^{t_f} \left[x(t) + z(t) - s(t) \right] dt, \tag{7}$$

which represents the concentration of $CD4^+$ T and CTL cells and the cost measured in terms of chemotherapy strength and a combination of duration and intensity. The optimal control problem consists in determining a control function $s \in L^1\left([0, t_f], \mathbb{R}\right)$ that maximizes the cost functional (7) subject to the control system (4), initial conditions (5) and control constraints (6). This optimal control problem is solved analytically and numerically.

Acknowledgements

This research was supported by the Portuguese Foundation for Science and Technology (FCT) within projects UID/MAT/04106/2013 (CIDMA) and PTDC/EEI-AUT/2933/2014 (TOCCATA). Rocha is also supported by the Ph.D. fellowship SFRH/BD/107889/2015; Silva by the post-doc fellowship SFRH/BPD/72061/2010.

References

[1] R. Culshaw, S. Ruan, R. Spiteri, *Optimal HIV treatment by maximising immune response*, J. Math. Biol., 48 (2004), 545–562.

[2] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[3] K. Hattaf, N. Yousfi, *Optimal Control of a Delayed HIV Infection Model with Immune Response Using an Efficient Numerical Method*, ISRN Biomathematics, (2012), 1–7.

[4] A. Kaddar, A. Abta and H. T. Alaoui, A comparison of delayed SIR and SEIR epidemic models, Nonlinear Anal. Model. Control **16** (2011), no. 2, 181–190.

[5] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, San Diego, 1993.

[6] D. Li and W. Ma, Asymptotic properties of a HIV-1 infection model with time delay, J. Math. Anal. Appl. **335** (2007), no. 1, 683–691.

[7] J. E. Mittler, B. Sulzer, A. U. Neumann and A. S. Perelson, *Influence of delayed viral production on viral dynamics in HIV-1 infected patients*, Math. Biosci., 152 (1998), 143–163.

[8] P. W. Nelson, J. D. Murray and A. S. Perelson, *A model of HIV-1 pathogenesis that includes an intracellular delay*, Mathematical Biosciences, 163 (2000), 201–215

[9] C. J. Silva and D. F. M. Torres, A TB-HIV/AIDS coinfection model and optimal control treatment, Discrete Contin. Dyn. Syst. **35** (2015), no. 9, 4639–4663.

[10] D. Wodarz, M. A. Nowak, Specific therapy regimes could lead to long-term immunological control of HIV, Proc. Natl. Acad. Sci. **96** (1999), 14464–14469.

[11] H. Zhu and X. Zou, *Dynamics of a HIV-1 infection model with cell-mediated immune response and intracellular delay*, Discrete and Continuous Dynamical System, Series B, 12 (2009), 511–524.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Analyzing criminal networks using Formal Concept Analysis with negative attributes

Rodriguez-Jimenez, J.M.[1], Cordero, P.[1], Enciso, M.[1] and Mora, A.[1]

[1] *Andalucia Tech, University of Malaga*

emails: jmrodriguez@ctima.uma.es, pcordero@uma.es, enciso@lcc.uma.es, amora@ctima.uma.es

Abstract

Development of methods that use negative attributes in Formal Concept Analysis is a new area that explores the power of negative information in data sets. These methods increase the knowledge given by classical methods and enable researchers to make faster and better decisions. One of this important contribution in the real world is the research about criminal organisations by police forces. There exist some processes that reduce and points to the needed information for investigating and stamping criminal networks.

Key words: formal concept analysis, negative attributes, policing, criminal networks

1 Introduction

Since data mining explore the use of negative attributes in basket analysis, many efforts are dedicated to explore this area for obtaining best results for knowledge discovering. If a customer usually buys an itemset of products, but avoid getting other itemset, for marketing purposes, this information can be analyzed to distribute products in markets or to predict how new products will be accepted or not.

But basket analysis is only a common example for data mining. There exist in different areas where the information given by negative attributes could increase the knowledge that we have and, related to this knowledge, the benefits that we can obtain in the real world.

Formal Concept Analysis explores special datasets where information can be stored as a binary data. In these datasets the presence of an attribute in an object is marked, since the absence of this attribute does not. Null values are not admitted. These datasets are stored in low size files that is an advantage to operate with them.

 ISBN: 978-84-608-6082-2

Classical Formal Concept Analysis methods only consider positive information avoiding the negative one. In most of cases, the presence of negative or absence of information is bigger than the positive one, so there is a deficiency in the use of the methods.

Police reports contain a description of different successes. This description covers different situations but does not have a fixed structure. However, there are some data that always appear when a person or a vehicle is reported: the registration number. This code is unique for each person or vehicle and have a fixed format so it is easy to be detected in a report with a simple text mining software.

2 Preliminaries

2.1 Formal Concept Analysis

In this section, the basic notions related with Formal Concept Analysis (FCA) [17] and attribute implications are briefly presented. See [4] for a more detailed explanation. A *formal context* is a triple $\mathbb{K} = \langle G, M, I\rangle$ where G and M are finite non-empty sets and $I \subseteq G \times M$ is a binary relation. The elements in G are named objects, the elements in M attributes and $\langle g, m\rangle \in I$ means that the object g has the attribute m. From this triple, two mappings $\uparrow$: $2^G \to 2^M$ and $\downarrow$: $2^M \to 2^G$, named derivation operators, are defined as follows: for any $X \subseteq G$ and $Y \subseteq M$,

$$X^{\uparrow} = \{m \in M \mid \langle g, m\rangle \in I \text{ for all } g \in X\} \tag{1}$$

$$Y^{\downarrow} = \{g \in G \mid \langle g, m\rangle \in I \text{ for all } m \in Y\} \tag{2}$$

$X^{\uparrow}$ is the subset of all attributes shared by all the objects in X and $Y^{\downarrow}$ is the subset of all objects that have the attributes in Y. The pair $(\uparrow, \downarrow)$ constitutes a Galois connection between 2^G and 2^M and, therefore, both compositions are closure operators.

A pair of subsets $\langle X, Y\rangle$ with $X \subseteq G$ and $Y \subseteq M$ such that $X^{\uparrow} = Y$ and $Y^{\downarrow} = X$ is named a *formal concept* where X is its *extent* and Y its *intent*. These extents and intents coincide with closed sets wrt the closure operators because $X^{\uparrow\downarrow} = X$ and $Y^{\downarrow\uparrow} = Y$. Thus, the set of all formal concepts is a lattice, named *concept lattice*, with the relation

$$\langle X_1, Y_1\rangle \leq \langle X_2, Y_2\rangle \text{ if and only if } X_1 \subseteq X_2 \text{ (or equivalently, } Y_2 \subseteq Y_1) \tag{3}$$

This concept lattice will be denoted by $\mathfrak{B}(G, M, I)$.

The concept lattice can be characterized in terms of *attribute implications* being expressions $A \to B$ where $A, B \subseteq M$. An implication $A \to B$ holds in a context $\mathbb{K}$ if $A^{\downarrow} \subseteq B^{\downarrow}$. That is, any object that has all the attributes in A has also all the attributes in B. It is well known that the sets of attribute implications that are valid in a context satisfies the so called Armstrong's Axioms:

I	a	b	c	d	e
o_1		×	×		×
o_2	×	×			
o_3		×	×		×
o_4			×	×	

Table 1: A formal context

`[Ref]` Reflexivity: If $B \subseteq A$ then $\vdash A \to B$.

`[Augm]` Augmentation: $A \to B \vdash A \cup C \to B \cup C$.

`[Trans]` Transitivity: $A \to B, B \to C \vdash A \to C$.

A set of implications Σ is considered an *implicational system for* $\mathbb{K}$ if: an implication holds in $\mathbb{K}$ if and only if it can be inferred, by using Armstrong's Axioms, from Σ.

Armstrong's axioms allow us to define the closure of attribute sets wrt an implicational system (the closure of a set A is usually denoted as A^+) and it is well-known that closed sets coincide with intents. On the other hand, several kind of implicational systems has been defined in the literature being the most used the so-called Duquenne-Guigues (or stem) basis [6]. This basis satisfies that its cardinality is minimum among all the implicational systems and can be obtained from a context by using the renowned NextClosure Algorithm [4].

2.2 Negatives attributes

As we have mentioned in the introduction, classical FCA only discover knowledge limited to positive attributes in the context, but it does not consider information relative to the absence of properties (attributes). Thus, the Duquenne-Guigues basis obtained from Table 1 is $\{e \to bc, d \to c, bc \to e, a \to b\}$. Moreover, the implications $b \to c$ and $b \to d$ do not hold in Table 1 and therefore they can not be derived from the basis by using the inference system. Nevertheless, both implications correspond with different situations. In the first case, some objects have attributes b and c (e.g. objects o_1 and o_3) whereas another objects (e.g. o_2) have the attribute b and do not have c. On the other side, in the second case, any object that has the attribute b does not have the attribute d.

A more general framework is necessary to deal with this kind of information. In [16], we have tackled this issue focusing on the problem of mining implication with positive and negative attributes from formal contexts. As a conclusion of that work we emphasized the necessity of a full development of an algebraic framework that was initiated in [15].

First, we begin with the introduction of an extended notation that allows us to consider the negation of attributes. From now on, the set of attributes is denoted by M, and its

$I \cup \overline{I}$	a	b	c	d	e	$\overline{a}$	$\overline{b}$	$\overline{c}$	$\overline{d}$	$\overline{e}$
o_1		×	×		×	×			×	
o_2	×	×						×	×	×
o_3		×	×		×	×			×	
o_4			×	×		×	×			×

Table 2: The formal context $(\mathbb{K}|\overline{\mathbb{K}})$

elements by the letter m, possibly with subindexes. That is, the lowercase character m is reserved for positive attributes. We use $\overline{m}$ to denote the negation of the attribute m and $\overline{M}$ to denote the set $\{\overline{m} \mid m \in M\}$ whose elements will be named negative attributes.

Arbitrary elements in $M \cup \overline{M}$ are going to be denoted by the first letters in the alphabet: a, b, c, etc. and $\overline{a}$ denotes the opposite of a. That is, the symbol a could represent a positive or a negative attribute and, if $a = m \in M$ then $\overline{a} = \overline{m}$ and if $a = \overline{m} \in \overline{M}$ then $\overline{a} = m$.

Capital letters A, B, C,... denote subsets of $M \cup \overline{M}$. If $A \subseteq M \cup \overline{M}$, then $\overline{A}$ denotes the set of the opposite of attributes $\{\overline{a} \mid a \in A\}$ and the following sets are defined:

- $\mathrm{Pos}(A) = \{m \in M \mid m \in A\}$
- $\mathrm{Neg}(A) = \{m \in M \mid \overline{m} \in A\}$
- $\mathrm{Tot}(A) = \mathrm{Pos}(A) \cup \mathrm{Neg}(A)$

Note that $\mathrm{Pos}(A), \mathrm{Neg}(A), \mathrm{Tot}(A) \subseteq M$.

Once we have introduced the notation, we are going to summarize some results concerning the mining of knowledge from contexts in terms of implications with negative and positive attributes [16]. A trivial approach could be obtained by adding new columns to the context with the opposite of the attributes [5]. That is, given a context $\mathbb{K} = \langle G, M, I\rangle$, a new context $(\mathbb{K}|\overline{\mathbb{K}}) = \langle G, M \cup \overline{M}, I \cup \overline{I}\rangle$ is considered, where $\overline{I} = \{\langle g, \overline{m}\rangle \mid g \in G, m \in M, \langle g, m\rangle \notin I\}$. For example, if $\mathbb{K}$ is the context depicted in Table 1, the context $(\mathbb{K}|\overline{\mathbb{K}})$ is those presented in Table 2. Obviously, the classical framework and its corresponding machinery can be used to manage the new context and, in this (direct) way, negative attributes are considered. However, this rough approach induces a non trivial growth of the formal context and, consequently, algorithms have a worse performance.

In our opinion, a deeper study was done by R. Missaoui et al. in [8] where an evolved approach has been provided. For first time –as far as we know– inference rules for the management of positive and negative attributes are introduced [9]. The authors also developed new methods to mine mixed attribute implications by means of the key notion [10].

In [16], we have developed a method to mine mixed implications whose main goal has been to avoid the management of the large $(\mathbb{K}|\overline{\mathbb{K}})$ contexts, so that the performance of the corresponding method has a controlled cost.

We extend the definitions of derivation operators, formal concept and attribute implication.

Definition 1 *Let $\mathbb{K} = \langle G, M, I\rangle$ be a formal context. We define the operators $\Uparrow$: $2^G \to 2^{M\cup\overline{M}}$ and $\Downarrow$: $2^{M\cup\overline{M}} \to 2^G$ as follows: for $X \subseteq G$ and $Y \subseteq M \cup \overline{M}$,*

$$\begin{aligned} X^{\Uparrow} = & \{m \in M \mid \langle g, m\rangle \in I \text{ for all } g \in X\} \\ \cup & \{\overline{m} \in \overline{M} \mid \langle g, m\rangle \notin I \text{ for all } g \in X\} \end{aligned} \tag{4}$$

$$\begin{aligned} Y^{\Downarrow} = & \{g \in G \mid \langle g, m\rangle \in I \text{ for all } m \in Y\} \\ \cap & \{g \in G \mid \langle g, m\rangle \notin I \text{ for all } \overline{m} \in Y\} \end{aligned} \tag{5}$$

Definition 2 *Let $\mathbb{K} = \langle G, M, I\rangle$ be a formal context. A* mixed formal concept *in $\mathbb{K}$ is a pair of subsets $\langle X, Y\rangle$ with $X \subseteq G$ and $Y \subseteq M \cup \overline{M}$ such $X^{\Uparrow} = Y$ and $Y^{\Downarrow} = X$.*

Definition 3 *Let $\mathbb{K} = \langle G, M, I\rangle$ be a formal context and let $A, B \subseteq M \cup \overline{M}$, the context $\mathbb{K}$ satisfies a* mixed attribute implication *$A \to B$, denoted by $\mathbb{K} \models A \to B$, if $A^{\Downarrow} \subseteq B^{\Downarrow}$.*

For example, in Table 1, as we previously mentioned, two different situations were presented. Thus, in this new framework we have that $\mathbb{K} \not\models b \to d$ and $\mathbb{K} \models b \to \overline{d}$ whereas $\mathbb{K} \not\models b \to c$ either $\mathbb{K} \not\models b \to \overline{c}$.

Once we have the implicational system with negative attributes, we can explore all the posible implications with the set of inference rules built by supplementing Armstrong's axioms with the following ones, introduced in [9]: let $a, b \in M \cup \overline{M}$ and $A \subseteq M \cup \overline{M}$,

`[Cont]` Contradiction: $\vdash a\overline{a} \to M\overline{M}$.

`[Rft]` Reflection: $Aa \to b \vdash A\overline{b} \to \overline{a}$.

3 Formal Concept Analysis in Police investigations

In Formal Concept Analysis, there exist some previous works that analyze criminal activities in different areas. Most of them are published by Poelmans and Elzinga, that is resumed in [11] how works their research.

- Radicalisation and terrorism were investigated by the National Police Service Agency of the Netherlands [1] that developed a model to classify (potential) jihadists. The goal of this model is to detect the potential jihadist to prevent him or her to enter the dangerous phase. They use Temporal Concept Analysis to visualize how a possible jihadist radicalizes over time.

- Domestic violence is one of the problems that is difficult to be solved due to lack of information about potential victims that do not report these situations. In [13, 14] the authors use Emergent Self Organizing Maps with Formal Concept Analysis to analyze different reports to locate potential victims. Text mining from police reports is used and show that there exist problems with labelling, confusing situations, missing values, etc.

- Human trafficking and forced prostitution research [12] try to discover in police reports using text mining these situations to filter out interesting persons for further investigation and used the temporal variant of Formal Concept Analysis to create a visual profile of these persons, their evolution over time and their social environment. For these purposes, finding in reports different specified indicators (loverboys, big amount of money, expensive cars, etc.) let researchers to obtain a lattice where suspects and victims can be related.

- Areas of greater intensity, called "hot spots", indicate where large amounts of reports were collected. With the help of geolocalization tools this research [7, 3] supports the distribution of resources, such as police officers, patrolling cars, and surveillance cameras, as well as the definition of strategies for crime combat and prevention.

- Other applications of Formal Concept Analysis is the analysis of pedosexual chat conversations [2] to prevent child abuse and violence.

4 Analyzing criminal networks.

Criminal networks are groups that work in illegal activities and have an extended structure that is very difficult to discover, for example, as we have mentioned in the previous section, that groups which are related to human trafficking. Relations between members are difficult to be discovered and, when there appears people that are not members of the group, the investigation grows and sometimes, police spend a valuable time following an erroneous path.

Police reports take notice of different events that contains important and specific information that are considered for this work as attributes. Works from Poelmans and Elzinga use text mining for extracting information but, in our research, we only use specific data easy to find in the reports. These attributes are persons and vehicles. Each person can be associated with a unique document reference and each vehicle to its registration number. Places are auxiliary information that we have to tread carefully because their values have to be grouped by zones, i. e., we have to consider that in reports sometimes appear the name of a business establishment or the street where is located, so for our initial method we are going to discard this information.

Due to different reasons, police officers often do not know all the information of previous reports from other police officers so the information have to be treated by staff dedicated to monitoring and analyzing these reports.

If we use a dataset where appears, in a temporal series, the itemsets that happen together, we can generate a concept lattice that represents the possible criminal network.

Using lattices with positive concepts, we can detect and check the possible hierarchy between members, but using a lattice with mixed attributes concepts, we can detect possible suspects that we can discard until we have more information, or that are not related between them.

Each concept in the lattice is going to be named with their attributes. In the upper zone of the lattice are located these concepts whose itemsets are the intersection of others. Pointing to itemsets that are located on the upper levels of the lattice we can investigate the most relevant attributes of the context.

Example 1 *There are 9 reports in Table 3 from police units that identify different groups of suspects that are located together near specific zones associated with criminal activities.*

Date	a	b	c	d	e	f	g	h	i	j	k
1	X			X					X		
2		X	X		X					X	
3	X							X	X		
4						X	X				X
5	X			X					X		X
6	X			X				X	X		
7						X	X				
8	X			X				X			
9							X				X

Table 3: Example of reports for criminal data

With this data set we can generate the concept lattice associated with this information. We can observe that the knowledge given by this lattice with only 9 objects it is not clear, so a bigger dataset will be complex and difficult to interpret.

Concepts with attributes a, g and k are located in the upper zone in the lattice that means that these attributes appear in common in several objects but not always have to be together. If we use NextClosure*[4], we can obtain a basis of implications from this context,* $\mathbb{B} = \{j \to bce, i \to a, h \to a, f \to g, e \to bcj, d \to a, c \to bej, b \to cej, bcejk \to adfghi, bcegj \to adfhik, ak \to di, ag \to bcdefhijk, adhik \to bcefgj, abcej \to dfghik\}$

If we check a, g and k separately, the knowledge that we have with positive attributes cannot offer any kind of information about our queries, so this basis is not useful for us. We can use the algorithm proposed in [16] to extract the mixed implicational system for obtaining

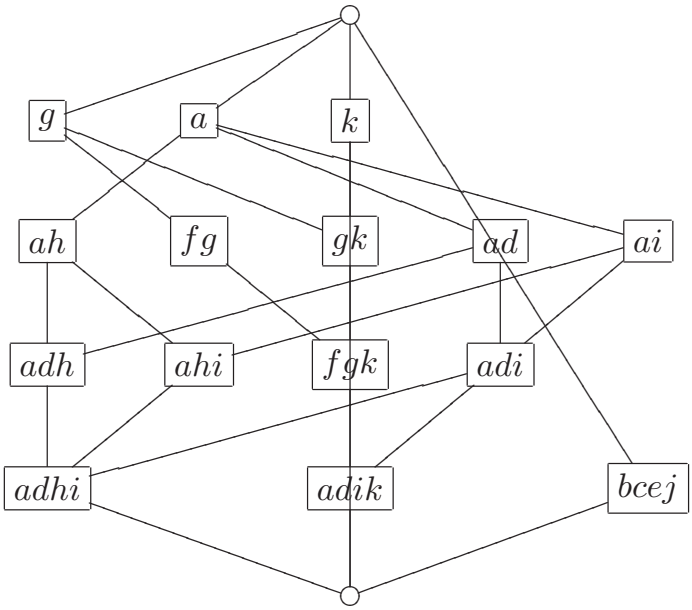


Figure 1: Concept lattice associated with Table 3.

additional knowledge. This implication system not only contains more implications, but due to inference rules that use negative attributes, offers more useful knowledge. $\Sigma = \{k \to \overline{b}\overline{c}\overline{e}\overline{h}\overline{j}, \overline{j} \to \overline{b}\overline{c}\overline{e}, j\overline{k} \to \overline{a}bc\overline{d}e\overline{f}\overline{g}\overline{h}\overline{i}, i \to a\overline{b}\overline{c}e\overline{f}\overline{g}\overline{j}, \overline{h}\overline{i} \to \overline{a}\overline{d}, \overline{g} \to \overline{f}, g \to \overline{a}\overline{b}\overline{c}\overline{d}\overline{e}\overline{h}\overline{i}\overline{j}, \overline{f}\overline{k} \to \overline{g}, \overline{f}\overline{g}\overline{i} \to \overline{k}, \overline{e} \to \overline{b}\overline{c}\overline{j}, d \to a\overline{b}\overline{c}\overline{e}\overline{f}\overline{g}\overline{j}, \overline{d}\overline{i} \to \overline{a}\overline{h}, \overline{d}\overline{h} \to \overline{a}\overline{i}, \overline{d}\overline{f}\overline{g} \to \overline{k}, \overline{c} \to \overline{b}\overline{e}\overline{j}, \overline{b} \to \overline{c}\overline{e}\overline{j}, \overline{b}\overline{c}\overline{e}h\overline{j}\overline{k} \to a\overline{f}\overline{g}, \overline{b}\overline{c}\overline{e}\overline{f}\overline{g}\overline{j} \to a, a \to \overline{b}\overline{c}\overline{e}\overline{f}\overline{g}\overline{j}, \overline{a}\overline{d}\overline{f}\overline{g}\overline{h}\overline{i}\overline{k} \to bcej, a\overline{b}\overline{c}\overline{d}\overline{e}\overline{h}\overline{i}\overline{j} \to g\}$

Now the queries return to us information using mixed implications 1, 7, and 19 that say which attributes are not related with them. Attributes $bcej$ *are not related with attributes* a, g *and* k*;* f *and* g *are not related with* a, d, h *and* i*. We can observe that there are 3 not related itemsets* $adhi$*,* fg *and* $bcej$ *and the attribute* k *that connects the 2 first itemsets. We can adapt the context to show this situation (see Table 4).*

If we discard object 2*, we can simplify the lattice and have a clear structure of the criminal network and the relation between attributes.*

How to interpret the lattice is so important to take into account the roles of the components of the network. In the lattice we cannot deduce a hierarchy structure for the criminal network, but the itemsets located in the concepts at the top of the lattice points to members that are interesting to be investigated specifically.

Date	a	d	h	i	k	f	g	b	c	e	j
1	X	X		X							
3	X		X	X							
5	X	X		X	X						
6	X	X	X	X							
8	X	X	X								
4					X	X	X				
7						X	X				
9					X		X				
2								X	X	X	X

Table 4: Example of reports for criminal data with relocated attributes

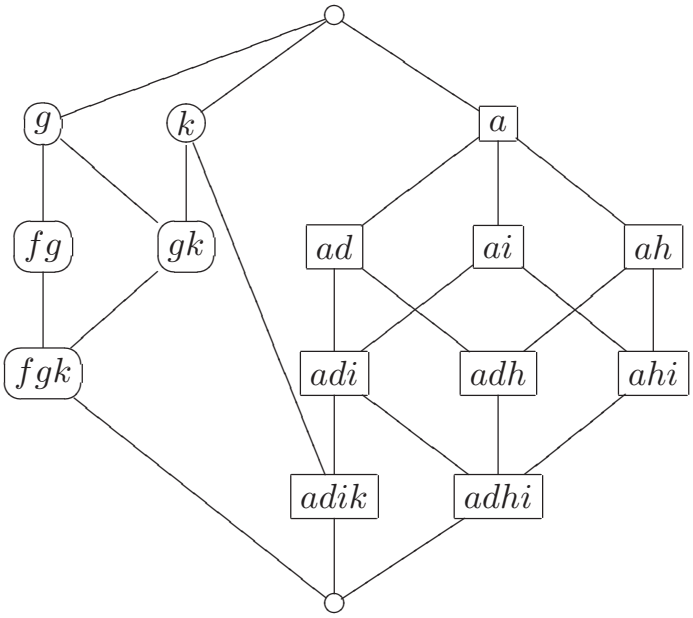


Figure 2: Concept lattice with cleared relations.

5 Conclusions and future works

In this work we use the previous theory developed about negative attributes in Formal Concept Analysis in a real world case to explore and analyze criminal networks.

Since each Police department has its own report database with different formats, we cannot propose a specific algorithm that detects the identification number for a person or the registration number for a vehicle, but since this data have a detailed format, it is easy

to develop software that recognizes the information in each report.

In the shown example, we can observe how our method works and how it simplifies the study of criminal networks for police forces.

We are studying different applications to existing problems in the real world that have been previously studied with Formal Concept Analysis, but only with the positive information. The additional knowledge could open a door to new results that can be applied in data mining or in artificial intelligence areas.

6 Acknowledgements

Supported by grant TIN2014-59471-P of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

[1] P. Elzinga, J. Poelmans, S. Viaene, G. Dedene, and S. Morsing. Terrorist threat assessment with formal concept analysis. In *IEEE International Conference on Intelligence and Security Informatics, ISI 2010, Vancouver, BC, Canada, May 23-26, 2010, Proceedings*, pages 77–82, 2010.

[2] P. Elzinga, K. E. Wolff, and J. Poelmans. Analyzing chat conversations of pedophiles with temporal relational semantic systems. In *2012 European Intelligence and Security Informatics Conference, EISIC 2012, Odense, Denmark, August 22-24, 2012*, pages 242–249, 2012.

[3] A.M.G. Farias, M.E. Cintra, A. F. Castro, and D.C. Lopes. Criminal hot spot detection using formal concept analysis and clustering algorithms.

[4] B. Ganter. Two basic algorithms in concept analysis. *Technische Hochschule, Darmstadt*, 1984.

[5] G. Gasmi, S. Ben Yahia, E. M. Nguifo, and S. Bouker. Extraction of association rules based on literalsets. In *DaWaK*, pages 293–302, 2007.

[6] J.L. Guigues and V. Duquenne. Familles minimales d implications informatives resultant d un tableau de donnees binaires. *Mathematiques et Sciences Sociales*, 95:5–18, 1986.

[7] Q-A. Kester. Visualization and analysis of geographical crime patterns using formal concept analysis. *CoRR*, abs/1307.8112, 2013.

[8] R. Missaoui, L. Nourine, and Y. Renaud. Generating positive and negative exact rules using formal concept analysis: Problems and solutions. In *ICFCA*, pages 169–181, 2008.

[9] R. Missaoui, L. Nourine, and Y. Renaud. An inference system for exhaustive generation of mixed and purely negative implications from purely positive ones. In *CLA*, pages 271–282, 2010.

[10] R. Missaoui, L. Nourine, and Y. Renaud. Computing implications with negation from a formal context. *Fundam. Inform.*, 115(4):357–375, 2012.

[11] J. Poelmans, P. Elzinga, and G. Dedene. Retrieval of criminal trajectories with an fca-based approach, 2013.

[12] J. Poelmans, P. Elzinga, G. Dedene, S. Viaene, and S. O. Kuznetsov. A concept discovery approach for fighting human trafficking and forced prostitution. In *Conceptual Structures for Discovering Knowledge - 19th International Conference on Conceptual Structures, ICCS 2011, Derby, UK, July 25-29, 2011. Proceedings*, pages 201–214, 2011.

[13] J. Poelmans, P. Elzinga, S. Viaene, and G. Dedene. A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. In *Advances in Data Mining. Applications and Theoretical Aspects, 9th Industrial Conference, ICDM 2009, Leipzig, Germany, July 20-22, 2009. Proceedings*, pages 247–260, 2009.

[14] J. Poelmans, P. Elzinga, S. Viaene, and G. Dedene. Curbing domestic violence: instantiating C-K theory with formal concept analysis and emergent self-organizing maps. *Int. Syst. in Accounting, Finance and Management*, 17(3-4):167–191, 2010.

[15] J. M. Rodríguez-Jiménez, P. Cordero, M. Enciso, and A. Mora. A generalized framework to consider positive and negative attributes in formal concept analysis. In *CLA*, pages 267–278, 2014.

[16] J. M. Rodríguez-Jiménez, P. Cordero, M. Enciso, and A. Mora. Negative attributes and implications in formal concept analysis. *Procedia Computer Science*, 31(0):758 – 765, 2014. 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.

[17] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In *Rival, I. (ed.): Ordered Sets*, pages 445–470. Boston, 1982.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A Rendezvous Framework for the Automatic Deployment of Services in Cluster Computing

Cristina Rodríguez-Quintana[1], Antonio F. Díaz[1], Julio Ortega[1], Raúl H. Palacios [1] and Andrés Ortiz[2]

[1] *Department of Architecture and Computer Technology, University of Granada*

[2] *Department of Communications Engineering, University of Málaga*

emails: crodriguez@ugr.es, afdiaz@ugr.es, jortega@ugr.es, raulhp@ugr.es, aortiz@ic.uma.es

Abstract

High-performance computing (HPC) systems are increasing their resources allowing the creation of complex services. Dynamic configuration management tools are needed to carry out the deployments of modern distributed services; however, specific solutions are implemented in every case. In this paper, we present a framework for dynamic configuration and management where servers rendezvous with clients and share their information automatically. The internal mechanisms evaluate overall requeriments and undertake actions to maintain the services robustly working. Although we have designed it for general use, it is oriented to distributed filesystems, such as AbFS. Through the evaluation, the proposed model takes low overhead and simplify the creation of new dynamic services.

Key words: Scalability, High-Availability, Filesystems, Consul

1 Introduction

An important issue to create robust distributed services is to management dynamic configurations and define mechanisms which maintain the main function of the global system under fault-tolerant conditions. Some automatic processes are needed to perform some tasks in a cluster computing environment with hundreds of nodes where servers or even network can fail.

Although manual or hardcoded configurations are easy to implement, the final system can provoke numerous problems such us: scalabiliy, resilient to failures, reduced visibility

 ISBN: 978-84-608-6082-2

and auditability, difficult configuration changes or even the generation of multiple points of failure.

There are several key issues to manage distributed resources and services: dynamic configuration, fault tolerant capabilities, health checking, and leader election. An ideal service should consider some of these elements. Some algorithms, protocols and tools have been developed to solve specific problems.

Paxos [7] is an algorithm that solves distributed consensus while tolerating host failures, network partitions, and message loss. Paxos allows multiple hosts to reach consensus on a single value by relying on majority decisions. Raft [9] is another consensus algorithm for managing a replicated log. It separates the key elements of consensus, such as leader election, log replication, and safety, and it enforces a stronger degree of coherency to reduce the number of states that must be considered. Chubby [1] is a fault-tolerant lock service. It has been designed to run on a small number of hosts, where each node holds a replica of a simple database. It is used in The Google File System [3] to choose primary nodes which ensure data consistency.

Zookeeper [5] is a high-performance coordination service for distributed applications written in Java. It maintains strong consistent, based on Zab protocol (Paxos-like) and quorum. All dataset must fit in memory. It implements a shared hierarchical namespace, with ephemeral node support and access control list to each node. Consul [4] is a tool for discovering and configuring services. It is written in Go, and based on Gossip protocol for all the nodes and consensus protocol (Raft-based) for servers. It also supports access control list, kev/value storage and multi datacenter capabilities.

Fault tolerant and redundant elements are needed to guarantee reliability systems; however, some studies related to proactive fault tolerance [8] can improve overall functioning. Futhermore, it can be applied to storage systems [6]. Even though these algorithms are intented to cover specific needs, specific solutions are implemented in every case. In this work we present a general framework to simplified the most common elements.

2 RendezVous Approach

We describe a model that allows dynamic system configuration and fault-tolerant based on agents that offer different services in a reliable manner. These services are executed transparently by agents which communicate to each other in a safe way.

Futhermore, the proposed model includes distributed virtual objects that simplify the execution of global and local code. These objects do not have to be associated with a particular host and have properties and methods that can run redundantly. Even though a host fails, another one can continue with the execution.

The framework provides the following resources:

- Dynamic configuration: Once a configuration is created, it must allow modify by

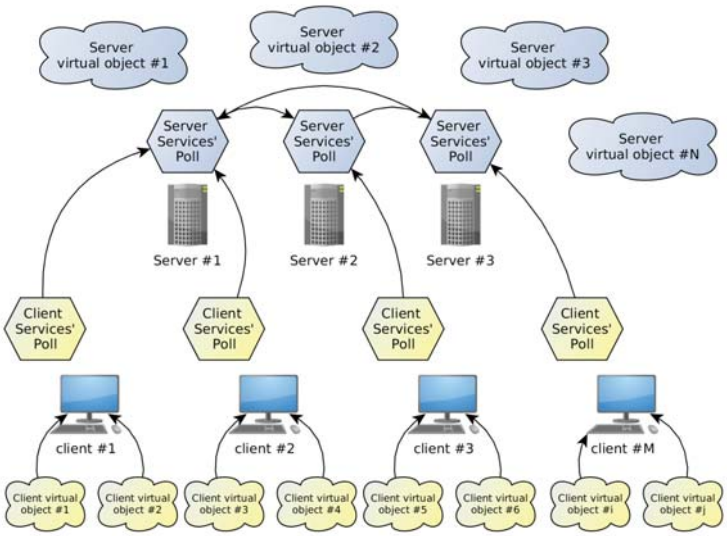


Figure 1: Rendezvous framework interconexion diagram.

adding or removing resources. This is also necessary to provide scalability.

- Runtime changes: Non-stop services require to operate without disruption.
- Consistent information: Despite changes and server failures, all nodes must have the same shared information.
- Health check: It is a critical component that prevent using services that are unhealthy.
- Dynamic Global administration: There are some nodes that must monitor the overall operation of the system.
- Leader election: Some resources can be shared to balance workload.
- Fault tolerant operation: If a resource fails the system can continue working.
- Resynchronization: If there is redundancy and a node fails, there are mechanisms to recover from a failure.
- How the software is updated: Software are continuosly evolving with the implementation of new funtionality and eliminating bugs, so automatic update mechanism can reduce failures during the process.

- How preventively operation is monitored: Preventing failures with the monitoring of several variables (e.g.S.M.A.R.T data in disks) detecting which devices can fail in a short period of time.

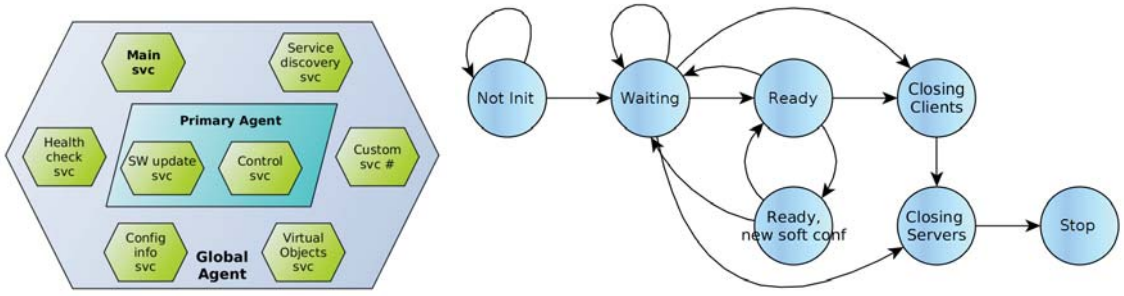
Although the framework can be used by any application, it is oriented to distributed filesystem, such as AbFS [2] with fault tolerant and scalability needs.

3 Implementation

The rendezvous framework was implemented in Python using some tool for discovering and configuring services such as Consul. It provides several key features: service discovery, health checking, key/value store and multi datacenter capabilities. The framework is composed of the following components:

- Admin servers: Responsible of the global operation.
- Servers: Offer resources and services to clients and can implement redundant elements.
- Clients: Access to servers.
- Services: Components that are shared.
- Virtual objects: can reside in any node and execute code without interruption.

Figure 1 shows the main components and how different services interact automatically sharing vital information to maintain the effective performance of the system.



(a) Services' Pool. (b) Predefined states in Virtual State Machine.

Figure 2: Some internal resources from Rendezvous Framework.

As shown in Figure 2.a, the system deploys two agents. A primary agent is the responsable of software updates and execute code remotely. Another global agent is in

charge of the rest of services: health checking, config information, service discovery (with leader election), virtual object environment and the possibility of setting up custom services.

Virtual objects can interact with each other and provides an hierarchy of objects. These objects can send and receive events asynchronously. There is a global event queue: This queue is redundant so when an event is generated it is replicated in the global queue so the event is not lost by an unexpected failure. If the node goes down, another node is responsible for continuing execution so the overall process is always alive. 3 to 5 servers are recommended to avoid failure scenarios. Virtual objects can include a state machine with a default set of states, as shown in Figure 2.b. These states can be extended as required.

4 Results

The framework has been tested in a cluster with 16 server nodes equipped with: 2.27GHz Intel(R) Xeon(R) E5520 CPUs, 16GB RAM memory, 1TB Local Disk and OS CentOS 6.6. Our scenario has 3 servers and 13 clients. Some times have been obtained:

Table 1: Execution times (in seconds) of several framework operations.

	Max	Average	Min	Std.Dev
Server write	0.072	0.042	0.037	0.01
Client access	0.026	0.005	0.002	0.007
Event response	0.05	0.047	0.045	0.0015
Health checking	0.0028	0.0027	0.0026	0.000082
Leader election	0.026	0.005	0.02	0.007
Update configuration	0.13	0.09	0.07	0.017
Virtual object creation	0.07	0.013	0.007	0.019

The system provides optimal response time considering that management processes do not require high speed operations.

5 Summary and future work

We presented a rendezvous framework that performs automatic resource management in HPC systems. It provides fault tolerant based on agents that offer different services in a reliable manner. These services are executed transparently by servers and clients agents to communicate each other in a safe way. The proposed model includes distributed virtual objects that simplify the execution of global and local code. This framework was developed in Python combining some tools for coordinate elements in a robust and dynamic way. It offers fast response time and simplifies dynamic configuration in distributed systems. We want to improve the framework with new features and define a robust test bench.

Acknowledgements

This work has been partially supported by European Union FEDER and the Spanish Ministry of Economy and Competitiveness TIN2015-67020-P, FPA2015-65150-C3-3-P, and PROMEP/103.5/13/6475 UAEH-146.

References

[1] Michael Burrows. The chubby lock service for loosely-coupled distributed systems. In *7th Symposium on Operating Systems Design and Implementation (OSDI '06), November 6-8, Seattle, WA, USA*, pages 335–350, 2006.

[2] Antonio F. Díaz, Mancia Anguita, Hugo E. Camacho, Erik Nieto, and Julio Ortega. Two-level hash/table approach for metadata management in distributed file systems. *The Journal of Supercomputing*, 64(1):144–155, 2013.

[3] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles 2003, SOSP 2003, Bolton Landing, NY, USA, October 19-22, 2003*, pages 29–43, 2003.

[4] HashiCorp. Consul. HashiCorp, 2016.

[5] Patrick Hunt, Mahadev Konar, Flavio Paiva Junqueira, and Benjamin Reed. Zookeeper: Wait-free coordination for internet-scale systems. *USENIX Annual Technical Conference*, 8:9, 2010.

[6] Xinpu Ji, Yuxiang Ma, Rui Ma, Peng Li, Jingwei Ma, Gang Wang, Xiaoguang Liu, and Zhongwei Li. *Algorithms and Architectures for Parallel Processing: 15th International Conference, ICA3PP 2015, Zhangjiajie, China, November 18-20, 2015, Proceedings, Part III*, chapter A Proactive Fault Tolerance Scheme for Large Scale Storage Systems, pages 337–350. Springer International Publishing, Cham, 2015.

[7] Leslie Lamport. Paxos Made Simple. *SIGACT News*, 32(4):51–58, December 2001.

[8] Antonina Litvinova, Christian Engelmann, and Stephen L. Scott. A proactive fault tolerance framework for high-performance computing. In *Proceedings of the 9th IASTED International Conference on Parallel and Distributed Computing and Networks, PDCN 2010*, pages 105–110, 2010.

[9] Diego Ongaro and John Ousterhout. In search of an understandable consensus algorithm. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference*, USENIX ATC'14, pages 305–320, Berkeley, CA, USA, 2014. USENIX Association.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Modelling the effects of a differentiated mortality by phenotypic traits on the genotypic distribution

Héctor Rojas-Castro[1] **and Fernando Córdova-Lepe**[1]

[1] *Facultad de Ciencias Básicas, Universidad Católica del Malule*

emails: hrojas@ucm.cl, fcordova@ucm.cl

Abstract

A closed, diploid and panmictic population, which is a harvestable resource having a Malthusian growth, is considered. According to their genotype and for a specific locus, it is compartmentalized in three subpopulations: two homozygotes and the heterozygote one. Under the assumption that the population is extracted by alternating closures with comparatively very short open seasons, the dynamic of the coupled genotypic frequencies of each subpopulation are modeled using an impulsive differential system. In addition, it is assumed that each genotype expresses a phenotypic characteristic that differentiates the mortality rates by catchability. Some results about the long-term genotypic distribution are obtained and presented as threshold conditions on the parameter space.

Key words: genetic variability, genotypic distribution, differentiated mortality, impulsive differential equations.

1 Introduction

Genetic variability and loss of biodiversity have become topics of interest to biologists and geneticists that focus and try to optimize their efforts on the implementation of conservation measures, see [1]. This assertion is supported, for instance, by the existence of a large number of related genetic diversity studies of specific populations, and the different conservation strategies. Examples of this type of studies can be seen in [3, 8, 9, 14].

Our work aims to study the genotypic distribution over time of a population which is a self-regenerative resource. We have in mind, for purposes of the concepts to be used, a fishery resource. It will be assumed that a portion of the population is captured periodically at relatively short time intervals, the open seasons. Genetic variability is registered by using

 ISBN: 978-84-608-6082-2

the genotypic frequencies and its behavior is studied over time, considering the exploitation of the resource as a selection process, in the evolutionary sense, see [2, 10, 11]. The above under the assumption that the extraction method can facilitate capture of the individuals who express certain phenotypic characteristics due to an specific genotype, thus favoring those individuals who are more difficult to capture or that they are discarded by the method used to carry out the extraction.

The fishery activity is a very good example of the situation that we seek to describe, because the methods or fishing gear used in the capture process can be more effectives in some individuals, more than others, who have certain characteristics, generally related to the size, see [16]. In [7, 15, 13], examples of different fish stocks that undergo changes in their genetic variability caused by the capture are shown.

For simplicity of the model, it will be assumed that the extraction process has two different time scales. A time interval without captures (e.g. a closure) where population increases in a natural way and, for other hand, a brief harvest period which is modeled as an instant, a pulse fishery see [6]. On the first scale, the population growth is modeled by a system of ordinary differential equations, where the variables are the genotypic frequencies. Since the captures have a very short duration, harvest instants, the changes in the abundace by fishery mortality are considered as pulses, this determines a second discrete system of equations. Thus, the dynamic for the genotype frequencies will be modeled by the coupling of the two systems of equations just mentioned, this is, by an impulsive system, see [12].

The paper is organized as follows: First, in section 2, the model is introduced. For this population abundances models and their respective frequencies are set, sections 2.1 and 2.2 respectively. Defined above, the impulsive model with differentiated capture described in section 2.3. Second, in Section 3 presents the main results.

2 The model

Let us consider a closed, diploid and panmictic population, which supports a Malthusian growth model. The letters A and a denote the two possible alleles for a given locus and it is assumed that there is not necessarily a dominance between them. Thus, it is possible to compartmentalize the population in three types of individuals by genotype for the locus under study: AA, Aa and aa. The functions $P_{AA}(t)$, $P_{Aa}(t)$ and $P_{aa}(t)$ represent the respective genotypic subpopulations at time t. So that, $P(t) = P_{AA}(t) + P_{Aa}(t) + P_{aa}(t)$ is the abundance of the total population.

Genotype frequencies, see [17], correspond to the relative proportion of existing genotypes in a population, this is, $P_{AA}(\cdot)/P(\cdot)$, $P_{Aa}(\cdot)/P(\cdot)$ and $P_{aa}(\cdot)/P(\cdot)$, which we denote respectively by $x(\cdot)$, $y(\cdot)$ and $z(\cdot)$. Note that $x(t) + y(t) + z(t) = 1$ for all t.

Similarly, see [17], *allele frequencies* is the ratio of a specific allele in the set of possible alleles for a given locus. If $f_A(t)$ and $f_a(t)$ are the allelic frequencies for alleles A and a

respectively at instant t, then it is possible to calculate them as:

$$f_A(t) = x(t) + \frac{1}{2}y(t) \quad \text{y} \quad f_a(t) = z(t) + \frac{1}{2}y(t), \quad t \geq 0. \tag{1}$$

Clearly f_A and f_a have values in $[0, 1]$ and their sum is the unity.

2.1 Abundance of genotypic subpopulations

Given a pair of parents in the population, individuals of their progeny can be part of any of the genotypic groups AA, Aa or aa, based on the Mendel's laws, see [17]. The following table summarizes a counting method that allows us to distribute the descendants in each one of the described groups, depending on the group of their parents:

Type	Apareos	AA	Aa	aa
$AA \times AA$	$(q/P)\, P_{AA}\, P_{AA}$	1	0	0
$AA \times A\,a$	$(q/P)\, P_{AA}\, P_{A\,a}$	1/2	1/2	0
$AA \times a\,a$	$(q/P)\, P_{AA}\, P_{a\,a}$	0	1	0
$A\,a \times AA$	$(q/P)\, P_{A\,a}\, P_{AA}$	1/2	1/2	0
$A\,a \times A\,a$	$(q/P)\, P_{A\,a}\, P_{A\,a}$	1/4	1/2	1/4
$A\,a \times a\,a$	$(q/P)\, P_{A\,a}\, P_{a\,a}$	0	1/2	1/2
$a\,a \times AA$	$(q/P)\, P_{a\,a}\, P_{AA}$	0	1	0
$a\,a \times A\,a$	$(q/P)\, P_{a\,a}\, P_{A\,a}$	0	1/2	1/2
$a\,a \times a\,a$	$(q/P)\, P_{a\,a}\, P_{a\,a}$	0	0	1

(2)

where $q = p/2$, with p the number of apareos in which an individual can participate.

Based on table (2), it is possible to calculate the number of apareos that generate offspring with a given genotype. We denote by h_{AA} (respectively h_{Aa} or h_{aa}) the number of apareos that generates offspring of type AA for any time (respectively Aa or aa). Then

$$h_{AA} = \frac{q}{P}\left(P_{AA}^2 + \frac{1}{2}P_{AA}P_{Aa} + \frac{1}{2}P_{AA}P_{Aa} + \frac{1}{4}P_{Aa}^2\right),$$

that using (1) can be written as:

$$h_{AA} = \frac{q}{P}\left(P_{AA} + \frac{1}{2}P_{Aa}\right)^2 = q f_A^2 P. \tag{3}$$

Similarly we obtain:

$$h_{Aa} = \frac{q}{P}2\left(P_{AA} + \frac{1}{2}P_{Aa}\right)\left(P_{aa} + \frac{1}{2}P_{Aa}\right) = 2q f_A f_a P \tag{4}$$

and

$$h_{aa} = \frac{q}{P}\left(P_{aa} + \frac{1}{2}P_{Aa}\right)^2 = q f_a^2 P. \tag{5}$$

 ISBN: 978-84-608-6082-2

Thus, if ν and μ are respectively the average number of offspring per mating and the natural mortality rate of the population, then the number of individuals incorporated per unit time according their genotype is:

$$\left\{\begin{array}{rcl} P'_{AA}(t) & = & \nu\, h_{AA}(t) - \mu P_{AA}(t), \\ P'_{A\,a}(t) & = & \nu\, h_{A\,a}(t) - \mu P_{A\,a}(t), \\ P'_{a\,a}(t) & = & \nu\, h_{a\,a}(t) - \mu P_{a\,a}(t). \end{array}\right. \tag{6}$$

Replacing (3), (4) and (5) in (6) and assuming $\alpha = \nu \cdot q$, we obtain the system of differential equations that follows:

$$\left\{\begin{array}{rcll} P'_{AA}(t) & = & \alpha\, f_A^2(t)P(t) & -\mu P_{AA}(t), \\ P'_{A\,a}(t) & = & 2\alpha\, f_A(t)f_a(t)P(t) & -\mu P_{A\,a}(t), \\ P'_{a\,a}(t) & = & \alpha\, f_a^2(t)P(t) & -\mu P_{a\,a}(t). \end{array}\right. \tag{7}$$

Note that from the system (7), we obtain $P' = P'_{AA} + P'_{Aa} + P'_{aa} = (\alpha - \mu)P$, which shows the Malthusian character of population abundance.

2.2 Frequency of genotypic subpopulations

We want to know how do the genotype frequencies of the population vary in time. Since $x(\cdot) = P_{AA}(\cdot)/P(\cdot)$, we have $P^2(t)x'(t)$ is equal to $P'_{AA}(t)P(t) - P_{AA}(t)P'(t)$, $t \geq 0$. Substituting P'_{AA} for the first equality in (7), we get it $x' = \alpha f_A^2 - \alpha x$. By proceeding analogously for $y(\cdot)$ and $z(\cdot)$, the following system is obtained:

$$\left\{\begin{array}{rcl} x'(t) & = & \alpha\,[f_A^2(t) - x(t)], \\ y'(t) & = & \alpha\,[2f_A(t)f_a(t) - y(t)], \\ z'(t) & = & \alpha\,[f_a^2(t) - z(t)]. \end{array}\right. \tag{8}$$

where $f_A(\cdot)$ and $f_a(\cdot)$, as functions of $x(\cdot)$, $y(\cdot)$ and $z(\cdot)$, are given by (1).

2.3 Pulse fishery. Harvest differentiated by genotype

As it was said, the population is considered a resource for human consumption. A portion of it is captured at an uniformly distributed sequence of instants, for example, a fish population whose extraction occurs instantaneously between consecutives closures of equal duration. It is supposed that the number of individuals captured per unit time is different for each of the genotypic subpopulations P_{AA}, P_{Aa} and P_{aa}. We are under the assumption that each genotype expresses a phenotypic characteristic that facilitates or hinders the captures. Some examples can be found in [16].

To set the model we will assume the following hypotheses:

(a) *Pulse fishery*: We consider that catches occur in an increasing sequence of instants $\{t_k\}$ equally spaced, *i.e.*, $t_{k+1} - t_k = \tau$, for some $\tau > 0$, for all $k \geq 0$.

The catch times are defined in this way, for instance, by conservation reasons. We need introduce closures for allowing the resource extracted can dispose a time interval to recover the abundance. The open seasons can be considerably more shorter than the closures. Some real fisheries, that adjust to this model, are cited in [5, 6].

(b) *Schaefer Hypothesis*: In each genotypic subpopulation, with differentiated catchability, we assume that the extractaction per unit of biomass is proportional to the fishing effort exerted in every instant of capture. The capture effort is denoted by the parameter E. It is an indicator of the inputs of the production process, see [4].

Based on these hypotheses, we obtain that in the k–th extraction the number of individuals of each genotypic subpopulation immediately after (this is, in the limit to t_k by the right) is given by:

$$\left\{\begin{array}{rcl} P_{AA}(t_k^+) & = & (1-q_1E)P_{AA}(t_k), \\ P_{Aa}(t_k^+) & = & (1-q_2E)P_{A\,a}(t_k), \\ P_{aa}(t_k^+) & = & (1-q_3E)P_{a\,a}(t_k), \end{array}\right. \tag{9}$$

where the coefficients q_i, $i \in \{1,2,3\}$, are interpreted as a measure of catchability (determined by the phenotypic characteristics that manifests each genotype) of individuals with genotype AA, Aa and aa respectively. Thus, $q_1EP_{AA}(t_k)$ corresponds to the number of individuals belonging to the subpopulation P_{AA} that are captured at the instant t_k, so that $P_{AA}(t_k^+) = P_{AA}(t_k) - q_1EP_{AA}(t_k) = (1-q_1E)P_{AA}(t_k)$. Similarly we obtain expressions for $P_{Aa}(t_k^+)$ and $P_{aa}(t_k^+)$. Note also that from the above we have $P(t_k^+) = (1-q_1E)P_{AA}(t_k) + (1-q_2E)P_{Aa}(t_k) + (1-q_3E)P_{aa}(t_k)$. When $q = q_1 = q_2 = q_3$, we have $P(t_k^+) = (1-qE)\,P(t_k)$, the standard Schaefer Hypotheses.

Now, introducing the change of variables $x(\cdot) = P_{AA}(\cdot)/P(\cdot)$, $y(\cdot) = P_{Aa}(\cdot)/P(\cdot)$ and $z(\cdot) = P_{aa}(\cdot)/P(\cdot)$, we obtain:

$$\left\{\begin{array}{l} x(t_k^+) = Q_1\,x(t_k)/\{Q_1x(t_k) + Q_2y(t_k) + Q_3z(t_k)\}, \\ y(t_k^+) = Q_2\,y(t_k)/\{Q_1x(t_k) + Q_2y(t_k) + Q_3z(t_k)\}, \\ z(t_k^+) = Q_3\,z(t_k)/\{Q_1x(t_k) + Q_2y(t_k) + Q_3z(t_k)\}, \end{array}\right. \tag{10}$$

where $Q_i = 1 - q_iE$, $i \in \{1,2,3\}$.

These expressions represent the genotype frequencies for each genotype of interest at time t_k^+, that is, just after the removal process. After this instant the dynamics is defined by the ordinary system (8), until the instant t_{k+1}, where we have the capture process again.

The pair of equations (7)-(9) and (8)-(10) are examples of impulsive differential equations at fixed impulse times (IDE-FT).

 ISBN: 978-84-608-6082-2

3 Main Results

A technical tool, used to study impulsive systems, is to consider one of the stroboscopic associated discrete map. From the dynamics of this map, It is possible to analyze and interpret the dynamics of the original impulsive system. In order to understand the effect that the captures causes in the system (8)-(10), we want to find a stroboscopic mapping linking the value of a solution $(x(\cdot), y(\cdot), z(\cdot))$, in an instant t_k, with its value at the successor time t_{k+1}. For the above, it is necessary to consider the following preliminary result:

Lema: The solution $(x(t), y(t), z(t))$, $t \geq 0$, of (8), with initial condition $t = 0$ equal to (x_0, y_0, z_0) it is given by:

$$\left\{\begin{array}{rcrcl} x(t) & = & f_A^2(1-e^{-\alpha t}) & + & x_0 e^{-\alpha t}, \\ y(t) & = & 2f_A f_a(1-e^{-\alpha t}) & + & y_0 e^{-\alpha t}, \\ z(t) & = & f_a^2(1-e^{-\alpha t}) & + & z_0 e^{-\alpha t}. \end{array}\right. \tag{11}$$

Proof: Notice that (8) corresponds to a differential system with constant parameters. So that, its solution is immediate. Then by (11), for $t \in (t_k, t_{k+1}]$, we have:

$$x(t) = \left(x(t_k^+) + \frac{1}{2}y(t_k^+)\right)^2 \left(1 - e^{-\alpha(t-t_k)}\right) + x(t_k^+)e^{-\alpha(t-t_k)}. \tag{12}$$

Evaluating at $t = t_{k+1}$ and identifying $u_j = u(t_j)$, $u_j^+ = u(t_j^+)$, for all $j \geq 0$ and $u \in \{x, y, z\}$, we obtain:

$$x_{k+1} = \left(x_k^+ + \frac{1}{2}y_k^+\right)^2 \left(1 - e^{-\alpha\,\tau}\right) + x_k^+ e^{-\alpha\,\tau}. \tag{13}$$

Considering expressions as (13) for variables y and z, and pulses given in (10), we obtain:

$$\left\{\begin{array}{rcl} x_{k+1}\,\sigma_k^2 & = & (Q_1 x_k + \frac{1}{2}Q_2 y_k)^2(1-e^{-\alpha\,\tau}) + Q_1 x_k \sigma_k e^{-\alpha\,\tau}, \\ y_{k+1}\,\sigma_k^2 & = & \left\{\begin{array}{l} 2(Q_1 x_k + \frac{1}{2}Q_2 y_k)(Q_3 z_k + \frac{1}{2}Q_2 y_k)(1-e^{-\alpha\,\tau}) \\ +Q_2 y_k \sigma_k e^{-\alpha\,\tau} \end{array}\right. \\ z_{k+1}\,\sigma_k^2 & = & (Q_3 z_k + \frac{1}{2}Q_2 y_k)^2(1-e^{-\alpha\,\tau}) + Q_3 z_k \sigma_k e^{-\alpha\,\tau} \end{array}\right. \tag{14}$$

where $\sigma_k = Q_1 x_k + Q_2 y_k + Q_3 z_k$, $k \geq 0$.

To study the dynamics determined by (14) we consider the following cases:

3.1 Equal catchability for all genotypic subpopulations

Suppose the catchability of a homozygous subpopulation is different from the catchability of the other genotypes. Without loss of generality, we can assume that the subpopulation

of genotype aa (homozygous) is different from the other two subpopulations (the other case is analogous). Thus, the system (14) can be presented as:

$$\left\{\begin{array}{lcl} x_{k+1} & = & \left\{\frac{x_k+\frac{1}{2}y_k}{x_k+y_k+\eta z_k}\right\}^2(1-p)+\left\{\frac{x_k}{x_k+y_k+\eta z_k}\right\}p, \\ \\ y_{k+1} & = & 2\left\{\frac{(x_k+\frac{1}{2}y_k)(\eta z_k+\frac{1}{2}y_k)}{(x_k+y_k+\eta z_k)^2}\right\}(1-p)+\left\{\frac{y_k}{x_k+y_k+\eta z_k}\right\}p, \\ \\ z_{k+1} & = & \left\{\frac{\eta z_k+\frac{1}{2}y_k}{x_k+y_k+\eta z_k}\right\}^2(1-p)+\left\{\frac{\eta z_k}{x_k+y_k+\eta z_k}\right\}p, \end{array}\right. \tag{15}$$

where $\eta = Q_3/Q_1$ and $p = e^{-\alpha\,\tau}$.

Using that $y = 1-x-z$, the system (15) can be rewritten with a smaller dimension

$$\left\{\begin{array}{lcl} x_{k+1} & = & F(x_k, z_k), \\ z_{k+1} & = & G(x_k, z_k), \end{array}\right. \tag{16}$$

where $F, G : \Delta \to \Delta$ are defined by:

$$\left\{\begin{array}{lcl} F(x,z) & = & \frac{1}{4}\left(\frac{1+x-z}{1+(\eta-1)z}\right)^2(1-p)+\left(\frac{x}{1+(\eta-1)z}\right)p, \\ \\ G(x,z) & = & \frac{1}{4}\left(\frac{1-x+(2\eta-1)z}{1+(\eta-1)z}\right)^2(1-p)+\left(\frac{\eta z}{1+(\eta-1)z}\right)p, \end{array}\right. \tag{17}$$

with $\Delta = \{(x,z)/x \geq 0, z \geq 0 \wedge x+z \leq 1\}$.

The mapping is well defined:
Let $(x,z) \in \Delta$, then clearly $F(x,z), G(x,z) \geq 0$. In addition, we note that: $F(x,z) + G(x,z) = 1-\frac{1}{2}\left(\frac{1+x-z}{1+(\eta-1)z}\right)\left(\frac{1-x+(2\eta-1)z}{1+(\eta-1)z}\right)(1-p)+\left(\frac{1-x-z}{1+(\eta-1)z}\right)p < 1$. Therefore, the mapping is well defined.

Equilibrium states:
In order to obtain fixed points, We solve the system $x = F(x,z)$ and $z = G(x,z)$. Thus, fixed points of the system (16) are:

$$P_x = (1,0),\ \ P_z = (0,1)\ \text{ and }\ P = \left(\frac{p-\eta}{\eta-1}, \frac{p-1}{\eta-1}\right).$$

The point P is discarded because it is not in the domain Δ of definition of the discrete mapping (14). Then, if the system trajectories converge towards a singularity, this should be P_x or P_z.

Theorem 1: Let us suppose that $q_1 = q_2$ and $q_3 \neq q_1$. Then, if $k \to \infty$, the trajectories $\{(x_k, y_k, z_k)\}$ of (14) are convergent to:

(a) State $(1,0,0)$ if $q_1 < q_3$, i.e. the least capturable homozygous subpopulation dominates.

(b) State $(0,0,1)$ if $q_1 > q_3$, i.e. the least capturable homozygous subpopulation dominates.

Proof: Note that for each point $P(x^*, z^*) \in \Delta$, you can associate the straight line:

$$L : z = x + b(x^*, z^*), \text{ where } b(x^*, z^*) = (z^* - x^*) \tag{18}$$

that is, the line containing the point P and with unitary slope.

Given a trajectory $\{(x_k, z_k)\}_{k\geq 0}$ defined by the mapping (16), we have:

(i) If $\eta < 1$, then $b_k = b(x_k, z_k) > b(x_{k+1}, z_{k+1}) = b_{k+1}$

(ii) If $\eta > 1$, then $b_k = b(x_k, z_k) < b(x_{k+1}, z_{k+1}) = b_{k+1}$

In either cases, the line $L_{(x_{k+1},z_{k+1})}$ is parallel to $L_{(x_k,z_k)}$ with a lower intercept if $\eta < 1$ and a greater intercept if $\eta > 1$. Indeed, given $(x_k, y_k) \in \Delta$, then:

$$L_{(x_k,z_k)} : z = x - (x_k - z_k), \tag{19}$$

Subtracting the equations of (16) we obtain:

$$L_{(x_{k+1},z_{k+1})} : z = x - \left(\frac{x_k - \eta z_k}{1 + (\eta - 1)z_k}\right). \tag{20}$$

It is easy to prove that $(x_k - z_k) < (x_{k+1} - \eta z_{k+1})/(1 + (\eta - 1)z_{k+1})$ (resp. $>$), if $\eta < 1$ (resp. $>$)

Case (i):
In this case, the succession of intercepts $\{b_k\}_{k\geq 0}$ is a strictly decreasing succession and also bounded below by $b(1,0) = -1$, therefore, it is convergent.

Let b_∞ the limit of the succession $\{b_k\}$. If $b_\infty > -1$, then accumulation points of the trajectory $\{(x_k, z_k)\}_{k\geq 0}$, defined by mapping (16), belong to the line segment $I_\infty = L_\infty \cap \Delta$, where $L_\infty : z = x + b_\infty$.

Let (x_a, z_a) some of these accumulation points. Then, by continuity of the mapping (F, G) defined in (16), the point (x_a^1, z_a^1), the iterated of (x_a, z_a), It is also an accumulation point. But if $(x_a^1, z_a^1) \neq (x_a, z_a)$, then $b(x_a^1, z_a^1) < b(x_a, z_a) = b_\infty$. Therefore $(x_a^1, z_a^1) \notin I_\infty$, which contradicts the fact that it is accumulation point, so that $(x_a^1, z_a^1) = (x_a, z_a)$, this is, (x_a, z_a) is a fixed point. Therefore $(x_a, z_a) = (1, 0)$.

Case (ii):
In a completely analogous way, it is proved that if $\eta > 1$, then the trajectory $\{x_k, z_k\}$ converges to $P_z(0,1)$. $\diamond$

 ISBN: 978-84-608-6082-2

Clearly, the above result is analogous for $q_3 = q_2$, $q_1 \neq q_3$. Furthermore, this result implies that under the conditions assumed loss of genetic variability is absolute, While watching the population one allele, resulting in a fully homozygous population, for the locus under study.

3.2 Differential catchability for heterozygous population

We suppose that the catchability of a heterozygous subpopulation is different from the catchability of other genotypes. This is, $q_1 = q_3$ and $q_1 \neq q_2$. Under these assumptions the system (14) can be written as:

$$\begin{cases} x_{k+1} &= \left\{\frac{x_k+\frac{\eta}{2}y_k}{x_k+\eta y_k+z_k}\right\}^2(1-p)+\left\{\frac{x_k}{x_k+\eta y_k+z_k}\right\}p, \\ y_{k+1} &= 2\left\{\frac{(x_k+\frac{\eta}{2}y_k)(z_k+\frac{\eta}{2}y_k)}{(x_k+\eta y_k+z_k)^2}\right\}(1-p)+\left\{\frac{y_k}{x_k+\eta y_k+z_k}\right\}p, \\ z_{k+1} &= \left\{\frac{z_k+\frac{\eta}{2}y_k}{x_k+\eta y_k+z_k}\right\}^2(1-p)+\left\{\frac{z_k}{x_k+\eta y_k+z_k}\right\}p. \end{cases} \tag{21}$$

where $\eta = Q_2/Q_1$ and $p = e^{-\alpha\,\tau}$.

Similarly to what was done in the previous section, the system (21) can be expressed as a two-dimensional discrete system, using that $y = 1 - x - z$:

$$\begin{cases} F(x,z) &= \left(\frac{x-\frac{1}{2}\eta}{1+(\eta-1)z}\right)^2(1-p)+\left(\frac{x}{1+(\eta-1)z}\right)p, \\ G(x,z) &= \frac{1}{4}\left(\frac{1+x-z}{1+(\eta-1)z}\right)^2(1-p)+\left(\frac{\eta z}{1+(\eta-1)z}\right)p, \end{cases} \tag{22}$$

with $\Delta = \{(x,z)/x \geq 0, z \geq 0 \wedge x + z \leq 1\}$.

Theorem 2: Suppose $q_1 = q_3$ and $q_2 \neq q_1$. Then the trajectories $\{(x_k, y_k, z_k)\}$ of the system (14) are convergent and

(a) If $q_2 > q_1$ and $x_0 > z_0$, they tend to $(1, 0, 0)$.

(b) If $q_2 > q_1$ and $x_0 < z_0$, they tend to $(0, 0, 1)$.

(c) If $q_2 > q_1$ and $x_0 = z_0$, they tend to P^*.

(d) If $q_2 < q_1$, they tend to P^* .

With $P^* = (x^*, 1 - x^* - z^*, z^*)$, where (x^*, z^*) is the only solution in Δ, of the system $x = F(x, z)$ and $z = G(x, z)$, which is different of $(1, 0)$ and $(0, 1)$.

 ISBN: 978-84-608-6082-2

Proof: The proof of **Theorem 2** is analogous (in the reasoning and its techniques) to the proof of **Theorem 1**. ⋄

In biological terms, the above result assures that if the catchability of the heterozygote subpopulation is greater than that of the other subpopulations, then it occurs a total loss of genetic variability, to the locus under study, except when initial genotypic frequencies are equal to subpopulations homozygous (uncommon in the biological context). Moreover, if the catchability of the heterozygote subpopulation is the smallest, then the genetic variability in the population remains, and the genotype frequencies converge, over time, to the values described by P^*.

Acknowledgements

This work has been partially supported by Univerisdad Católica del Maule, through the research project VRIP-UCM434166 and PMI-UCM1310.

References

[1] F. Allendorf and G, Luikart., *Conservation and the genetics of populations*, Blackwell Publishing, 2007.

[2] T. Beacham, *Variability in size and age at sexual maturity of American plaice and yellowtail flounder in the Canadian Maritimes Region of the northwest Atlantic Ocean*, Canadian Technical Report of Fisheries and Aquatic Sciences, 1983.

[3] E. Benevides, M. Vallinoto, A. Fetter, et al., *When physical oceanography meets population genetics: The case study of the genetic/evolutionary discontinuity in the endangered goliath grouper (Epinephelus itajara; Perciformes: Epinephelidae) with comments on the conservation of the species*, Biochemical Systematics and Ecology **56** (2014) 255–266.

[4] C. Clark., *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, John Wiley and Sons, 1990.

[5] F. Córdova and M. Pinto., *Mathematical Bioeconomics. Explotation of resources and preservation*, Cubo Mat. Educ. **239** (2002) 49–54.

[6] F. Córdova, R. Del Valle and Robledo, G, *A pulse fishery model with closures as function of the catch: Conditions for sustainability*, Mathematical Biosciences **239** (2012) 169–177.

[7] K. Hindar, N. Ryman and F. Utter, *Genetic effects of cultured fish on natural populations*, Canadian Journal of Fisheries and Aquatic Sciences **48** (1991) 945–957.

[8] R. Leary, F. Allendorf and S. Forbes, *Conservation genetics of bull trout in the Columbia and Klamath River drainages*, Conservation Biology **7** (1993) 856–865.

[9] X. Li, S. Zhang, Z. Yang, et al, *Conservation genetics and population diversity of Erigeron breviscapus (Asteraceae), an important Chinese herb*, Biochemical Systematics and Ecology **49** (2013) 156–166.

[10] W. Ricker, *Causes of the decrease in age and size of chinook salmon (Oncorhyncus tshawytscha)*, Canadian Technical Report of Fisheries and Aquatic Sciences, 1980..

[11] W. Ricker, *Size and age of British Columbia sockeye salmon (Oncorhynchus nerka) in relation to environmental factors and the fishery*, Canadian Technical Report of Fisheries and Aquatic Sciences, 1982

[12] A. Samoilenko, and N. Perestyuk, *Impulsive differential equations*, World Scientific series on Nonlinear Science, 1995.

[13] O. Sandlund and T. Naesje, *Impact of a pelagic gill-net fishery on the polymorphic whitefish (Coregonus lavaretus L.sl.) population in Lake Femund, Norway*, Fisheries Research **7** (1989) 85–97.

[14] C. Siler, A. Lira-Noriega, and R. Brown., *Conservation genetics of Australasian sailfin lizards: Flagship species threatened by coastal development and insufficient protected area coverage*, Biological Conservation **169** (2014) 100–108.

[15] R. Silliman, *Selective and unselective exploitation of experimental populations of Tilapia mossambica*, Fishery Bulletin **73** (1975) 495–507.

[16] P. Smith, *Genetic diversity of marine fisheries resources: possible impacts of fishing*, FAO Fisheries Technical Paper **17** (1982) 661–692.

[17] A. Templeton, *Population genetics and microevolutionary theory*, John Wiley & Sons, Inc, 2006

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Energy Consumption of Stencil-Based MPDATA Algorithm

Krzysztof Rojek[1], Maria Barreda[2], Enrique S. Quintana-Ortí[2] and Roman Wyrzykowski[1]

[1] *Institute of Computer & Information Sciences, Czestochowa Univ. of Technology, Czestochowa (Poland)*

[2] *Depto. de Ingeniería y Ciencia de Computadores, Univ. Jaume I, Castellón (Spain)*

emails: krojek@icis.pcz.pl, mvaya@uji.es, quintana@uji.es, roman@icis.pcz.pl

Abstract

We perform an experimental evaluation of the impact that voltage-frequency scaling and concurrency throttling exert on the energy consumption of the MPDATA algorithm, a key component of the multiscale fluid model EULAG. This analysis reveals that, for this particular algorithm, optimizing for performance is equivalent to optimizing for energy efficiency on a system equipped with a 6-core Intel Xeon CPU.

Key words: Energy Efficiency, Voltage-Frequency Scaling (VFS), MPDATA, Performance Analysis, Multi-threaded Parallelism

1 Introduction

Performance analysis has traditionally focused on optimizing computational throughput of applications and/or reducing their execution time. However, the end of Dennard scaling [2] has promoted energy as a holistic design principle in par with performance. As a result, a considerable amount of works have recently analyzed the interaction among temperature-power-time-energy for a variety of applications. In addition, these studies have targeted all sorts of current architectures, including multicore processors, graphics accelerators, the Intel Xeon Phi, and clusters assembled using any of these technologies.

One particular aspect that many of these past works address is the use of (dynamic) voltage-frequency scaling (VFS) [3], sometimes combined with (dynamic) concurrency throttling (CT) [1], as a means to reduce power dissipation and/or energy consumption. In this

paper we contribute towards this goal by studying the impact of VFS and CT on the execution of MPDATA, a key component that strongly dictates the computational performance as well as energy consumption of the multiscale fluid model EULAG [5, 4].

The rest of the paper is structured as follows. In Section 2 we offer a brief review of MPDATA, emphasizing its role in the framework of EULAG. Section 3 contains the main contribution of our paper, namely an experimental evaluation of MPDATA under different VFS/CT configurations, interleaved with a few concluding remarks.

2 The MPDATA Algorithm

The MPDATA algorithm belongs to the group of non-oscillatory forward-in-time algorithms and performs a sequence of stencil computations [5]. In this class of iterative algorithms, a single iteration requires to call one instance of the MPDATA routine, and the iterations (also referred to as time steps) are performed sequentially.

MPDATA solves continuity equation describing the advection of a non-diffusive quantity Ψ in a flow field [4], namely

$$\frac{\partial \Psi}{\partial t} \; + \; div(V\Psi) = 0, \tag{1}$$

where V is the velocity vector. The algorithm is positive defined and by appropriate flux correction can also be monotonic. This is a desirable property for advection of positive definite variables such as specific humidity, cloud water/ice, aerosol particles, gaseous substances, etc. The spatial discretization of MPDATA is based on finite difference approximations.

In the first sub-step, advection of a prognostic field Ψ is computed with the standard donor-cell approximation, ensuring the first order of accuracy only. In the subsequent time step, corrections are applied to make the scheme more accurate; i.e., second order in space and time. In the corrective sub-step, the donor-cell approximation is used with new anti-diffusive velocities based on the advected fields. The procedure can be repeated many times, however typically no significant improvements are observed after more than 2 corrections.

3 Experimental Results and Conclusions

All experiments were performed on 6-core Intel Xeon CPU E5-2620 processor, executing the MPDATA algorithm with a grid of size $256 \times 256 \times 256$ and 51 time steps. The energy measurements were obtained using the Intel RAPL interface for PKG and DRAM, as well as the PMLIB library to collect power samples from an internal wattmeter providing the energy consumption of entire server (including motherboard with all connected components).

Tables 1 and 2 respectively report the energy consumption for the PKG+DRAM and the entire server. The MPDATA algorithm is scalable across all 6 CPU cores. In these results we observe that increasing the CPU frequency decreases the energy consumption,

Table 1: MPDATA energy consumption (PKG+DRAM) for different number of cores and frequency levels.

Frequency (GHz)	Number of cores					
	1	2	3	4	5	6
1.2	81,688.2	44,765.3	32,442.7	26,296.8	22,710.7	20,586.6
1.3	76,643.2	42,214.0	30,722.7	24,926.5	21,510.9	19,590.7
1.4	72,517.9	40,399.5	29,236.1	24,031.3	20,665.8	18,733.1
1.5	69,572.2	38,417.4	28,030.7	23,164.3	19,843.2	17,924.4
1.6	66,445.5	37,129.5	27,017.8	22,266.9	19,228.1	17,501.6
1.7	63,914.7	35,830.4	26,162.1	21,700.4	18,610.7	17,156.3
1.8	61,808.6	34,666.2	25,556.1	21,148.9	18,197.5	16,633.2
1.9	60,376.9	33,927.4	24,838.4	20,730.1	17,823.3	16,360.6
2.0	59,189.2	33,180.0	24,510.2	20,462.4	17,607.7	16,092.2

independently of the number of cores. In addition, exploiting the algorithm's concurrency, by increasing the number of cores/threads, also helps reduce the energy consumption. In other words, for this particular processor and algorithm, optimizing for performance is equivalent to optimizing for energy efficiency.

We can also conclude that, from the perspective of energy efficiency, the number of cores is more significant than the CPU frequency, as the energy can be reduced by a factor about 3.9 by changing the former parameter but only by a factor 1.4 by tampering with the frequency. When we compare the results obtained using Intel RAPL with the results achieved using pmlib, we note that the second returns consumption rates that are about 1.4 times higher than those collected from Intel RAPL, due to the additional energy required to operate system components other than the CPU and DRAM.

Acknowledgements

The researchers from Universidad Jaime I (UJI) were supported by the CICYT project TIN2014-53495-R of MINECO and FEDER. The researchers from Czestochowa University of Technology were supported by the National Science Centre, Poland under grant no. UMO-2015/17/D/ST6/04059. This work was partially performed during a short term scientific missing (STSM) from Krzysztof Rojek to UJI supported by the EU COST IC1305.

Table 2: MPDATA energy consumption (server) for different number of cores and frequency levels.

Frequency (GHz)	Number of cores					
	1	2	3	4	5	6
1.2	115,723.0	63,608.1	45,996.9	37,645.1	32,363.7	29,096.3
1.3	109,317.0	59,652.2	43,044.8	35,645.8	30,568.9	27,579.8
1.4	103,304.0	56,528.2	41,815.6	33,912.8	29,110.4	26,278.6
1.5	98,140.0	53,847.0	40,465.0	32,603.3	27,854.3	25,181.7
1.6	94,504.3	51,974.5	38,672.6	31,419.4	26,839.5	24,246.5
1.7	90,467.9	51,176.4	37,040.9	30,555.2	25,956.0	23,615.1
1.8	87,220.4	49,359.0	35,932.3	29,702.3	25,308.5	22,985.3
1.9	84,570.8	48,423.4	34,980.1	28,846.0	24,730.0	22,442.2
2.0	82,593.9	46,644.1	34,412.9	28,344.9	24,218.5	22,116.8

References

[1] M. Curtis-Maury, F. Blagojevic, C.D. Antonopoulos, and D.S. Nikolopoulos. Prediction-based power-performance adaptation of multithreaded scientific codes. *Parallel and Distributed Systems, IEEE Transactions on*, 19(10):1396–1410, 2008.

[2] R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, and A.R. LeBlanc. Design of ion-implanted MOSFET's with very small physical dimensions. *Solid-State Circuits, IEEE Journal of*, 9(5):256–268, 1974.

[3] E.N. Elnozahy, Michael Kistler, and Ramakrishnan Rajamony. Energy-efficient server clusters. In *Power-Aware Computer Systems Second International Workshop, PACS 2002*, volume 2325 of *Lecture Notes in Computer Science (LNCS)*, pages 179–197, Cambridge, MA, USA, 2003. Springer-Verlag.

[4] Joseph M. Prusa, Piotr K. Smolarkiewicz, and Andrzej A. Wyszogrodzki. EULAG, a computational model for multiscale flows. *Computers & Fluids*, 37(9):1193–1207, 2008.

[5] Piotr K. Smolarkiewicz. Multidimensional positive definite advection transport algorithm: an overview. *International Journal for Numerical Methods in Fluids*, 50(10):1123–1144, 2006.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A fuzzy regression approach using Bernstein polynomials for the spreads and an application to a real Economic context

Antonio Francisco Roldán López de Hierro[1], Juan Martínez-Moreno[2], Concepción Aguilar Peña[3] and Concepción Roldán[4]

[1] *Department of Quantitative Methods for Economics and Business, University of Granada*

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

[3] *Department of Statistics and Operations Research, University of Jaén*

[4] *Department of Statistics and Operations Research, University of Granada*

emails: aroldan@ugr.es, jmmoreno@ujaen.es, caguilar@ujaen.es, iroldan@ugr.es

Abstract

In this study we show how to employ Bernstein polynomials in order to overcome the problem of non-negativeness when we determine functions that modelize the spreads of the objective function in fuzzy regression analysis. We illustrate the applicability and effectiveness of our methodology through the analysis of a real Economic context.

Key words: Fuzzy regression, Fuzzy number, Bernstein polynomials.

1 Introduction

Regression analysis is a useful statistical tool which has many applications in almost all scientific areas, including Engineering, Environmental Sciences, Finance and Economics, Medicine, Biology, Psychology, etc. In these days, there is an increasing interest in dealing with fuzzy data because, in real problems, experimental data can usually be affected by a certainty degree of imprecision or subjectivity.

Among previous methodologies about fuzzy regression, we are interested in those which interpret their optimal solutions as a regression function consisting in a model for centers and other models for spreads. These processes can handle a wider variety of models but a natural problem arises: the models for the spreads can only take non-negative values. At

 ISBN: 978-84-608-6082-2

this point starts a large discussion in the context of fuzzy regression analysis which is not solved for the moment.

The main aim of the present contribution is to provide the researcher with a viable and simple way for analyzing regression relationships when the problem of non-negativity spreads appears. In this paper we present a new fuzzy methodology based on the good properties that Bernstein polynomials verify.

2 Preliminaries

Let $\mathbb{R}$ be the set of all real numbers and let $\mathbb{I} = [0, 1]$ denote the unit interval. A *fuzzy set on* $\mathbb{R}$ is a map $\mathcal{A} : \mathbb{R} \to \mathbb{I}$. A *fuzzy number* (for short *FN*) on $\mathbb{R}$ is a fuzzy set $\mathcal{A}$ on $\mathbb{R}$ such that, for all $\alpha \in]0, 1]$, the *α-level set* (or *α-cut*) $\mathcal{A}_\alpha = \{x \in \mathbb{R} : \mathcal{A}(x) \geq \alpha\}$ is a non-empty, closed subinterval of $\mathbb{R}$. A *triangular fuzzy number* (for short, *TFN*) is a FN $\mathcal{A} = (a_1/a_2/a_3)$, where $a_1, a_2, a_3 \in \mathbb{R}$ (called the *corners* of $\mathcal{A}$), $a_1 < a_2 < a_3$, defined by:

$$\mathcal{A}(x) = \begin{cases} \dfrac{x - a_1}{a_2 - a_1}, & \text{if } a_1 < x \leq a_2, \\ \dfrac{a_3 - x}{a_3 - a_2}, & \text{if } a_2 < x < a_3, \\ 0, & \text{in any other case.} \end{cases}$$

Let $\mathcal{T}$ be the family of all TFNs on $\mathbb{R}$. The number $A^c = a_2$ is its *center* and its *spreads* are $A^\ell = a_2 - a_1 \geq 0$ and $A^r = a_3 - a_2 \geq 0$. The center and the spreads also determine the TFN since $a_1 = A^c - A^\ell$, $a_2 = A^c$ and $a_3 = A^c + A^r$. In this way, instead of $\mathcal{A} = (a_1/a_2/a_3)$, from now on, we will also use the representation of a TFN through its center and its spreads, that is, we will write $\mathcal{A} = \mathrm{Tri}(A^c, A^\ell, A^r)$.

Given a non-empty set D, any mapping $F : D \to \mathcal{T}$ can also be represented by

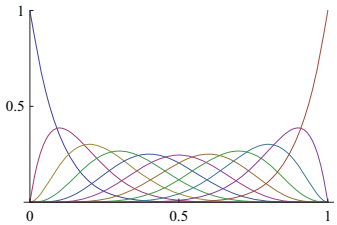
$$F = \mathrm{Tri}(F^c, F^\ell, F^r), \tag{1}$$

where $F^c : D \to \mathbb{R}$ and $F^\ell, F^r : D \to \mathbb{R}_0^+$ (only taking non-negative values) are the mappings $F^c(x) = F(x)^c$, $F^\ell(x) = F(x)^\ell$ and $F^r(x) = F(x)^r$ for all $x \in D$.

Bernstein polynomials are well-known from approximation theory. We recall here some of their properties. The *Bernstein polynomials of degree n* are defined as

$$b_{k,n}(x) = \binom{n}{k} x^k (1 - x)^{n-k},$$

where $n \in \mathbb{N} \cup \{0\}$ and $k \in \{0, 1, \ldots, n\}$.



For our purposes, 0^0 will mean 1. The most important property of Bernstein polynomials is that they only take non-negative values on $\mathbb{I}$, that is, $b_{k,n}(x) \geq 0$ for all $x \in \mathbb{I}$.

When $f : \mathbb{I} \to \mathbb{R}$ is a continuous function, its *Bernstein approach of degree* n is the polynomial

$$\sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_{k,n}(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k} \tag{2}$$

3 Bernstein estimation procedure

In this section we describe how to use Bernstein polynomials to estimate the relationship between two real RVs such that response variable only takes non-negative values.

Let (X, Y) be a bidimensional RV and assume that Y only takes non-negative values (that is, $Y \geq 0$ and $a_Y \geq 0$). Suppose that we are interested in analyzing the relationship between Y and X throughout an estimation mapping $\widehat{Y} = f(X)$. To do that, consider a random experiment in which we observe the variable (X, Y) on n statistical units, i.e., suppose that we have a random sample $\{(x_i, y_i)\}_{i=1}^n$ obtained from (X, Y). In particular, $y_i \geq 0$ for all $i \in \{1, 2, \ldots, n\}$.

To better explain the procedure to estimate Y from X, we distinguish whether the ranges of the variables are the unit interval $\mathbb{I}$ or not.

3.1 Unidimensional real random variables that take values in $\mathbb{I}$

Suppose that (X, Y) is a bidimensional RV such that both X and Y are $[0, 1]$-valued. Without loss of generality, we can suppose that $0 \leq x_1 < x_2 < \ldots < x_n \leq 1$. In this case, we will proceed as follows.

Step 1. Choose the natural number $N \in \mathbb{N}$ which we will use as the degree of the Bernstein polynomial regression function.

Step 2. Consider the piecewice linear function $\phi : \mathbb{I} \to \mathbb{I}$ that joins the points of the random sample.

Step 3. The Bernstein regression function will be

$$\widehat{Y}(x) = \sum_{k=0}^{N} \phi\left(\frac{k}{N}\right) b_{k,n}(x) = \sum_{k=0}^{N} \phi\left(\frac{k}{N}\right) \binom{N}{k} x^k (1-x)^{N-k} \quad \text{for all } x \in \mathbb{I}.$$

3.2 Unidimensional real random variables

If (X, Y) is a bidimensional RV valued on $[a_X, b_X] \times [a_Y, b_Y]$, we can consider a change of origin and scale to the interval $\mathbb{I}$. If $\widehat{Y}'_{x'}$ is the estimated Bernstein regression function for the transformed random sample, we finalize considering the regression mapping:

$$\widehat{Y}(x) = a_Y + (b_Y - a_Y)\, \widehat{Y}'\left(\frac{x - a_X}{b_X - a_X}\right) \quad \text{for all } x \in [a_X, b_X]\,.$$

Therefore $\widehat{Y}(x) \geq 0$ for all $x \in [a_X, b_X]$.

4 Bernstein estimation procedure applied to fuzzy regression theory

Let $X, \mathcal{Y}$ be two variables where X is a crisp RV and $\mathcal{Y}$ is a triangular FRV. Taking into account the canonical representation (1) of any mapping valued on $\mathcal{T}$, the regression model can be formalized as

$$\mathcal{Y} = \mathrm{Tri}(m(X, \underline{\omega}^c) + \varepsilon^c,\ s^{\,\ell}(X) + \varepsilon^\ell,\ s^{\,r}(X) + \varepsilon^r), \tag{3}$$

where $\varepsilon^c, \varepsilon_i^\ell$ and ε_i^s are the residuals (i.e., real-valued RVs such that $E[\varepsilon^c|X] = E[\varepsilon_i^\ell|X] = E[\varepsilon_i^r|X] = 0$ and whose variances are finite) and $\underline{\omega}^c = (\omega_1^c, \ldots, \omega_p^c)'$ is the $(p \times 1)$-vector of the parameters related to model for the center. To solve the regression problem, we are interested in obtaining a function $\widehat{\mathcal{Y}} : \mathrm{ran}\, X \to \mathcal{T}$ to predict $\mathcal{Y}$ from X. This function must be defined as

$$\widehat{\mathcal{Y}}_X = \mathrm{Tri}(\widehat{m}(X, \underline{\omega}^c),\ \widehat{s}^{\,\ell}(X),\ \widehat{s}^{\,r}(X)), \tag{4}$$

where $\widehat{m} : \mathrm{ran}\, X \to \mathbb{R}$ and $\widehat{s}^{\,\ell}, \widehat{s}^{\,r} : \mathrm{ran}\, X \to \mathbb{R}_0^+$ are arbitrary functions.

Consider a random sample $\{(X_j, \mathcal{Y}_j = \mathrm{Tri}(Y_j^c, Y_j^\ell, Y_j^r))\}_{j=1}^n$ obtained from $(X, \mathcal{Y})$. To estimate the mappings $\widehat{m}, \widehat{s}^{\,\ell}, \widehat{s}^{\,r}$ based on the random sample, center and spreads have to be treated differently.

- By considering the random sample $\{(X_j, Y_j^c)\}_{j=1}^n$, we may use any statistical package to fit a regression mapping $\widehat{m}$ considering the least square method.
- We propose to apply the Bernstein estimation procedure described in Section 3 to the random samples $\{(X_j, Y_j^\ell)\}_{j=1}^n$ and $\{(X_j, Y_j^r)\}_{j=1}^n$ to fit respective models for the spreads $\widehat{s}^{\,\ell}$ and $\widehat{s}^{\,r}$.

Finally, the fuzzy model obtained by this methodology is given in (4).

5 Application to a real Economic context

In 2010, Mosleh *et al.* [1] considered the following experimental study. Consumer reports uses a survey to collect data on the annual cost of repairs for 5 makes and models of automobiles. The fuzzy data are the fuzzy annual repair cost ($\mathcal{Y}$, in dollars) and the age of the automobile (X, in years). Data can be found in [1]. Notice that, in this case, spreads Y_j^ℓ and Y_j^r are not equal.

Age (years)	Cost (US dollars)			Reduced data to $[0,1]$		
X_j	Y_j^c	Y_j^ℓ	Y_j^r	X_j'	$(Y_j^\ell)'$	$(Y_j^r)'$
1	139	11	19	0.2	0.091667	0.158333
2	190	26	50	0.4	0.216667	0.416667
3	340	40	18	0.6	0.333333	0.15
4	358	18	42	0.8	0.15	0.35
5	470	70	115	1	0.583333	0.958333

Table 1: Transformed values.

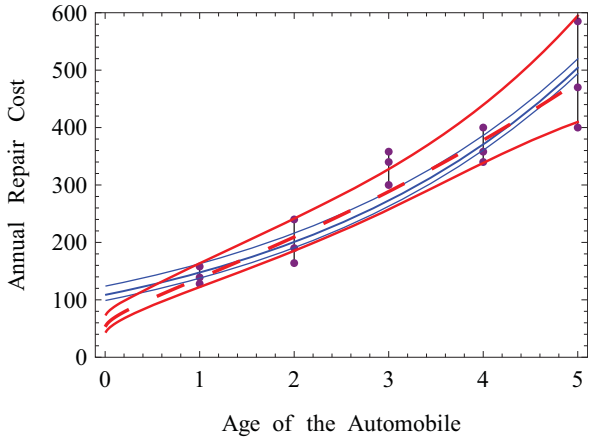


Figure 1: Comparison between Mosleh *et al.*'s model (in blue) and the proposed model (in red).

The proposed model in [1] is

$$\widehat{\mathcal{Y}}_{MOA}(x) = \mathrm{Tri}(108.6068, 9.8329, 15.1986)\, \mathrm{e}^{\mathrm{Tri}(0.3070,0{:}0062,0.0059)x}.$$

To apply our procedure, we consider that X takes values in $[0,5]$, Y^c in $[0,600]$ and the spreads are in $[0,120]$. Taking as degree $N=5$, the fuzzy model using Bernstein polynomials is $\widehat{\mathcal{Y}}_{\mathrm{BFM}}(x) = \mathrm{Tri}(\widehat{m}(x), \widehat{s}^\ell(x), \widehat{s}^r(x))$ where, for $x \in [0,5]$, the mappings $\widehat{m}(x)$, $\widehat{s}^\ell(x)$ and

$\widehat{s}^r(x)$ are given by:

$$\begin{aligned}
\widehat{m}(x) &= \exp\left(3.91764 + 1.00873\,\sqrt{x}\,\right),\\
\widehat{s}^{\ell}(x) &= 11 + 6\,x^2 - 1.28\,x^3 - 0.152\,x^4 + 0.05248\,x^5,\\
\widehat{s}^{r}(x) &= 19 + 12.4\,x^2 - 7.52\,x^3 + 1.704\,x^4 - 0.10848\,x^5.
\end{aligned}$$

Both models are plotted in Figure 1.

Acknowledgments

The first author is grateful to the Department of Quantitative Methods for Economics and Business of the University of Granada (Spain) for its economic support. This manuscript has been partially supported by Junta de Andalucía by projects FQM-268, FQM-178, FQM-245 and FQM-235 of the Andalusian CICYE.

References

[1] M. Mosleh, M. Otadi, S. Abbasbandy, *Evaluation of fuzzy regression models by fuzzy neural network*, J. Comput. Appl. Math. **234** (2010) 825–834.

[2] A.F. Roldán López de Hierro, J. Martínez-Moreno, C. Aguilar-Peña, C. Roldán López de Hierro, *A fuzzy regression approach using Bernstein polynomials for the spreads: Computational aspects and applications to economic models*, Math. Comput. Simulat. **128** (2016) 13–25.

[3] A.F. Roldán López de Hierro, J. Martínez-Moreno, C. Aguilar-Peña, C. Roldán, *Estimation of a fuzzy regression model using fuzzy distances*, IEEE Trans. Fuzzy Syst. **24** (2016) 344–359.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Symetry reductions and Conservation laws for a type of Fisher equations

M. Rosa[1] **and M.L Gandarias**[1]

[1] *Departamento de Matemáticas, University of Cádiz*

emails: maria.rosa@uca.es, marialuz.gandarias@uca.es

Abstract

Key words: Symmetry reductions, Fisher equation, reaction-diffusion, partial differential equations.

Generalizations of the Fisher equation are needed to more accutarely model complex diffusion and reactions effects found in many biological systems. There are many models that use nonlinear dispersal to describe the tendency for diffusion to increase due to overcrowding. Fisher equations are commonly used in biology for population dynamics models and in bacterial growth problems as well as development and growth of solid tumours. The theory of reaction-diffusion waves begins in the 1930s with the works in population dynamics, combustion theory and chemical kinetics. At the present time, it is a well developed area of research which includes qualitative properties of travelling waves for the scalar reaction-diffusion equation and for system of equations, complex nonlinear dynamics, numerous applications in physics, chemistry, biology, medicine [12]. Reaction-diffusion equations are conventionally used in chemical physics in order to describe concentration and temperature distributions. In this case, heat and mass transfer are described by the diffusion term while the reaction term describes the rate of heat and mass production.

The equation analyzed in this paper is a generalized Fisher equation

$$u_t = f(u) + \frac{1}{c(x)} \left(c(x)g(u)u_x\right)_x , \tag{1}$$

 ISBN: 978-84-608-6082-2

where $u(x,t)$ denotes the tumor cell density at location x and time t, being x and t the independent variables, $g(u)$ is the diffusion coefficient representing the active motility of cells depending on the variable u, $f(u)$ an arbitrary function and $c(x)$ an arbitrary function depending on the spatial variable x. In the particular case of $c(x) = 1$ and $g(u) = 1$, symmetry reductions and exact solutions were obtained using classical and nonclassical symmetries in [2]. When $c(x) = 1$ but $g(u)$ and $f(u)$ are arbitrary functions, equation

$$u_t = f(u) + (g(u)u_x)_x \tag{2}$$

is known as the density dependent diffusion-reaction equation which is mentioned by J.D. Murray [12] to model the advance of an advantageous gene through a geographic region. An exact solution of a quasilinear Fisher equation in cylindrical coordinates.

$$u_t = u(1-u) + \frac{1}{x}\left[xuu_x\right]_x, \tag{3}$$

by using Lie classical reductions was derived in terms of the Bessel functions in [1].
A complete classification of the classical symmetries and exact solutions of the Fisher equation in cylindrical coordinates and two arbitrary functions

$$u_t = f(u) + \frac{1}{x}\left(xg(u)u_x\right)_x, \tag{4}$$

were obtained in [10]. All the reductions were derived from the optimal system of subalgebras. Some of the reduced equations admitted Lie symmetries which yielded to further reductions. In [11], we have obtained Lie symmetries and conservation laws for the generalized Fisher equation with three arbitrary functions (1).

It is known that conservation laws play a significant role in the solution process of an equation or a system of differential equations. Although not all of the conservation laws of partial differential equations (PDEs) may have physical interpretation they are essential in studying the integrability of the PDEs. For variational problems, the Noether theorem [1] can be used for the derivation of conservation laws. For non variational problems there are different methods for the construction of conservation laws. In [6], a general theorem which does not require the existence of Lagrangians has been introduced by Ibragimov. This theorem is based on the concept of adjoint equations for nonlinear equations. The concept of strictly self-adjoint equations has been generalized [7, 3, 8]. After Ibragimov's results several papers appeared concerned with self-adjointness and its applications to PDEs [5] Recently, the property of nonlinear self-adjointness with differential substitution has been considered in [4].
In [9] we have determined the class of equations (1) which are nonlinearly self-adjoint. By using a general theorem on conservation laws proved by N.H. Ibragimov and the symmetry generators we found some conservation for some of these partial differential equations

without classical Lagrangians. The property of nonlinear self-adjointness with differential substitution has been considered in [4].

In Anco and Bluman gave a general treatment of a direct conservation law method for partial differential equations expressed in a standard Cauchy-Kovaleskaya form in particular for evolution equations

$$u_t = G(x, u, u_x, u_{xx}, \ldots, u_{nx}).$$

The nontrivial conservation laws are characterized by a multiplier λ with no dependence on u_t satisfying

$$\hat{E}[u]\left(\Lambda u_t - \Lambda G(x, u, u_x, u_{xx}, \ldots, u_{nx})\right) = 0. \tag{5}$$

Here

$$\hat{E}[u] := \frac{\partial}{\partial u} - D_t\frac{\partial}{\partial u_t} - D_x\frac{\partial}{\partial u_x} + D_x^2\frac{\partial}{\partial u_{xx}} + \ldots$$

is the Euler operator. The conservation law will be written

$$D_t(\Phi^t) + D_x(\Phi^x) \mid_{\Delta} = 0,$$

where Φ^t and Φ^x are called the conserved densities. The conserved current must satisfy

$$\Lambda = \hat{E}[u]\Phi^t \tag{6}$$

and the flux Φ^x is given by

$$\Phi^x = -D_x^{-1}(\Lambda G) - \frac{\partial \Phi^t}{\partial u_x}G + GD_x\left(\frac{\partial \Phi^t}{\partial u_{xx}}\right) + \ldots. \tag{7}$$

In [13], conditions for the existence of multipliers involving derivatives, whose existence is connected with the nonlinear self-adjointness with differential substitution has been given for evolution PDEs.

In this work, we will derive new nontrivial conservation laws for equation (1) by using the integral formulae of Anco and Bluman multiplier method. By using the symmetry generators and the optimal system of subalgebras previously derived in [11] we will consider

$$u_t = k_3 u + \frac{1}{k_1 e^{rx}}\left(k_1 e^{rx} k_2 u^{-2} u_x\right)_x. \tag{8}$$

We will perform the similarity reductions and we will search for exact solutions of physical and chemical interest. For equation (8) it happens that some of the associated systems admit symmetries that yield to potential symmetries of (8). These symmetries allow us to do new reductions of equation (8) which were unobtainable by using Lie classical symmetries.

Acknowledgements

The support of Junta de Andalucía group FQM-201 is gratefully acknowledged.

References

[1] A. H. Bokhari, M. T. Mustaf and F. D. Zaman, *An exact solution of a quasilinear Fisher equation in cylindrical coordinates*, Nonlinear Analysis **69** (2008) 4803–4805.

[2] P. A Clarkson, E. L Mansfield, *Symmetry reductions and exact solutions of a class of nonlinear heat equations*, *Physica D* **70** (1993) 250–288.

[3] M. L. Gandarias, *Weak self-adjoint differential equations*, J. Phys. A: Math. Theor. **44** (2011) 262001.

[4] M. L. Gandarias, *Nonlinear self-adjointness through differential substitutions*, Communications in Nonlinear Science and Numerical Simulation. **19** (2014) 3523–3528.

[5] M. L. Gandarias, M. S. Bruzón and M. Rosa, *Nonlinear self-adjointness and conservation laws for a generalized Fisher equation*, Communications in Nonlinear Science and Numerical Simulation **18** (2013) 1600–1606.

[6] N. H. Ibragimov, *Quasi self-adjoint differential equations*, Preprint Archives of ALGA **4** (2007) 55–60.

[7] N. H. Ibragimov, *The answer to the question put to me by L.V. Ovsyannikov 33 years ago*, Archives of ALGA **3** (2006) 80.

[8] N. H. Ibragimov, *Nonlinear self-adjointness and conservation laws*, Journal of Physics A: Mathematical and Theoretical **44** (2011) 432002.

[9] M. Rosa, M. S. Bruzón, M. L Gandarias., *A conservation law for a generalized chemical Fisher equation*, Journal of Mathematical Chemistry **53** (2014) 941–948.

[10] M. Rosa, M. S. Bruzón, M. L. Gandarias, *Symmetry analysis and exact solutions for a generalized Fisher equation in cylindrical coordinates*, Commun Nonlinear Sci Numer Simulat **25** (2015) 74–83.

[11] M. Rosa, M. S. Bruzón, M. L. Gandarias, *Lie Symmetry Analysis and Conservation Laws for a Fisher Equation with Variable Coefficients*, Appl. Math. Inf. Sci. **9** No. 6 (2015) 1–11.

[12] J. D. Murray, *Mathematical Biology*, Third Edition. Springer-Verlag New York Berlin Heidelberg (2002).

[13] Zhi-Yong Zhang, On the existence of conservation law multiplier for partial differential equations, *Communications in Nonlinear Science and Numerical Simulation* **20** (2015) 338-351.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An improved class of estimators of a linear parameter using auxiliary information in randomized response surveys

María del Mar Rueda[1] **and Beatriz Cobo**[1]

[1] *Department of Statistics and Operational Research, University of Granada, Spain*

emails: mrueda@ugr.es, beacr@ugr.es

Abstract

This work proposes a general class of estimators for a linear parameter of a sensitive variable using auxiliary information. Under a general randomized response model, the optimum estimator in this class is derived. Design based properties of proposed estimators are obtained. Several examples reflect the potential gains from the use of the proposed estimators instead of the customary estimators.

Key words: Auxiliary information, Randomized Response Technique, Horvitz-Thompson estimator, Sample surveys
MSC 2000: 62D05

1 Introduction

In many surveys it becomes necessary to probe into areas considered sensitive and potentially embarrassing. The validity of self-reports of sensitive attitudes and behaviors suffers from the tendency of individuals to distort their responses towards their perception of what is socially acceptable. As a consequence, studies self-report measures consistently underestimate the prevalence of undesirable attitudes or behaviors and overestimate the prevalence of desirable attitudes or behaviors. In an attempt to reduce this bias, [1] developed the randomized response technique (RRT). His idea spawned a vast volume of literature, see, for instance [2], [3], [4], [5].

Most research into RRT techniques deals exclusively with the interest variable and does not make explicit use of auxiliary variables in the construction of estimators. In this paper we suggest a class of estimators for a finite population total when the population totals of the auxiliary variables are known.

 ISBN: 978-84-608-6082-2

2 Estimating linear parameters in RRT

Consider a finite population U, consisting of N different individuals. Let y_i be the value of the sensitive aspect under study for the ith population element. Our aim is to estimate the finite population total $Y = \sum_{i=1}^{N} y_i$ of the variable of interest y or the population mean $\bar{Y} = 1/N \sum_{i=1}^{N} y_i$.

Assume that a sample s of individuals is chosen according to a general design p with inclusion probabilities $\pi_i = \sum_{s \ni i} p(s), i \in U$.

The interviews of individuals in the sample s are conducted in accordance with a RR model. Since y_i is not directly available from the respondent, y_i is estimated through the randomized response obtained from the ith respondent. Suppose that the ith respondent has to conduct a RR trial independently and z_i is the randomized response (or scrambled response) for the trial. For each $i \in s$ the RR induces a revised randomized response r_i that is an unbiased estimation of y_i with $\phi_i = V(r_i)$. Then an unbiased estimator for the population total of the sensitive characteristic y is given by

$$\hat{Y}(r) = \sum_{i \in s} \frac{r_i}{\pi_i}$$

The variance of $\hat{Y}(r)$ is given by

$$V(\hat{Y}(r)) = \left[\frac{1}{2} \sum_{i \neq} \sum_{j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \sum_{i \in U} \frac{\phi_i}{\pi_i} \right] = V_{ht} + \sum_{i \in U} \frac{\phi_i}{\pi_i}$$

and an unbiased estimator of $V(\hat{Y}(r))$ is

$$\hat{V}(\hat{Y}(r)) = \left[\frac{1}{2} \sum_{i \neq} \sum_{j \in s} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j} \right)^2 + \sum_{i \in s} \frac{\hat{\phi}_i}{\pi_i} \right] = \hat{V}_{ht} + \sum_{i \in s} \frac{\hat{\phi}_i}{\pi_i}$$

where π_{ij} are the second order inclusion probabilities of the design p. One assumes that the second order inclusion probabilities are non null.

3 Estimators in the presence of auxiliary information

3.1 A general class of estimators for the total

Proposed estimators consider k auxiliary variables $x_1, \ldots, x_k$, for which the population totals $X_1, \ldots, X_k$, are known. Our goal is to estimate the population parameter Y by using observations of the variables $y, x_1, \ldots, x_k$ in the sample s, and the known population values $X_1, \ldots, X_k$ associated with the auxiliary variables. We note by $\widehat{X}_h$ the Horvitz-Thompson estimator of the total X_h $(h = 1, ..., k)$.

 ISBN: 978-84-608-6082-2

Motivated by [6], we suggest the class of estimators of Y

$$\widehat{Y}_g^r = \{G(\hat{Y}(r), u_1, ...u_k)\}, \tag{1}$$

where $G(\cdot)$ is a function of $u_h = \widehat{X}_h/X_h$, continuous in a closed convex sub-space, $P \subset \mathbb{R}^{k+1}$, containing the point $(Y, 1, ..., 1) = (Y, \mathbf{1})$, and such that

(A1) $G(Y, \mathbf{1}) = Y$

(A2) $G_0'(Y, \mathbf{1}) = 1$ where $G_0'(Y, \mathbf{1})$ denoting the first partial derivative of $G(\cdot)$ with respect to $\hat{Y}(r)$, and

(A3) The first and second order partial derivatives of $G(\cdot)$ exist and are also continuous and bounded in P.

Theorem 3.1 *Any estimator into the class (1) is asymptotically unbiased for Y.*

Proof
By expanding G about the point $(Y_\beta, \mathbf{1})$ in a first order Taylor series, it is found that

$$\widehat{Y}_g^r = G(Y, \mathbf{1}) + (\hat{Y}(r) - Y)G_0'(Y, \mathbf{1}) + \sum_{h=1}^{k} G_h'|_{(Y_\beta, \mathbf{1})}(u_h - 1) + O(n^{-1}) \tag{2}$$

where G_h' denotes the first order partial derivative with respect to u_h.
By taking expectations on both sides in (2) we obtain

$$E[\widehat{Y}g^r] = Y + E[\hat{Y}(r)] - Y + \sum_{h=1}^{k} E(\widehat{X}_h - X_h)\frac{G_h'|_{(Y,\mathbf{1})}}{X_h}.$$

but $E[\hat{Y}(r)] = E_d E_R(\hat{Y}(r)) = Y$, $E[\widehat{X}_h] = E_d(\widehat{X}_h) = X_h$. Thus $E[\widehat{Y}g^r] = Y + O(n^{-1})$ so the bias is of order n^{-1}.

Theorem 3.2 *The bias of the proposed class of estimators is given by:*

$$B[\widehat{Y}g^r] = \sum_{h<t} \frac{Cov(\widehat{X}_h, \widehat{X}_t)}{X_h X_t} G_{ht}''|_{(Y,\mathbf{1})} + \frac{1}{2}\sum_{h=1}^{k} \frac{V(\widehat{X}_h)}{X_h^2} G_{hh}''|_{(Y,\mathbf{1})}$$

$$+\frac{1}{2}\frac{V(\hat{Y}(r))}{Y} G_{00}''|_{(Y,\mathbf{1})} + \frac{1}{2}\sum_{h=1}^{k} \frac{Cov(\widehat{X}_h, \widehat{Y}^r)}{X_h} G_{0h}''|_{(Y,\mathbf{1})}$$

 ISBN: 978-84-608-6082-2

Proof.

By expanding G about the point $(Y_\beta, \mathbf{1})$ in a second order Taylor series,

$$\widehat{Y}_g^r = Y + (\hat{Y}(r) - Y_\beta) + \sum_{h=1}^{k} G'_h|_{(Y,\mathbf{1})}(u_h - 1) +$$

$$\sum_{h<t}(u_h - 1)(u_t - 1)G''_{ht}|_{(Y,\mathbf{1})} + \frac{1}{2}\sum_{h=1}^{k}(u_h - 1)^2 G''_{hh}|_{(Y,\mathbf{1})} +$$

$$\frac{1}{2}\sum_{h=1}^{k}(u_h - 1)(\hat{Y}(r) - Y)G''_{0h}|_{(Y,\mathbf{1})} + \frac{1}{2}(\hat{Y}(r) - Y)^2 G''_{00}|_{(Y,\mathbf{1})} + O(n^{-2})$$

where G''_{hj} denote the second order partial derivative with respect to u_h and u_j, G''_{0h} is the second order partial derivative with respect to Y_β and u_h, and G''_{00} is second order partial derivative respect to Y.

Taking expectations in the the above second degree approximation we obtain the approximate bias (of order $O(n^{-2})$) of the proposed estimator.

It is well known that for a general sampling design d,

$$V(\widehat{X}_h) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_{hi}}{\pi_i} - \frac{x_{hj}}{\pi_j}\right)^2$$

and

$$Cov(\widehat{X}_h, \widehat{X}_t) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_{hi}}{\pi_i} - \frac{x_{tj}}{\pi_j}\right)^2$$

$V(\hat{Y}(r))$ is given in section 2. Thus we need to obtain the $Cov(\widehat{X}_h, \widehat{Y}^r)$. Using the covariance theorem we have:

$$Cov(\widehat{X}_h, \hat{Y}(r)) = E_d(cov_R(\widehat{X}_h, \hat{Y}(r)) + cov_d(E_R(\widehat{X}_h), E_R(\hat{Y}(r))) =$$

$$E_d(0) + cov_d\left(\widehat{X}_h, \sum_{i\in s}\frac{y_i}{\pi_i}\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_{hi}}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Theorem 3.3 *The variance of any estimator into the class (1) verifies (up to terms of order n^{-1}):*

$$V(\widehat{Y}_g^r) \geq V(\hat{Y}(r)) - \sigma'\mathbf{\Sigma}^{-1}\sigma,$$

where $\mathbf{\Sigma} = (a_{ht})_{(k\times k)}$ with $a_{hh} = V(\widehat{X}_h)$, $a_{ht} = Cov(\widehat{X}_h, \widehat{X}_t)$ and

$\sigma = (Cov(\hat{Y}(r), \widehat{X}_1)^r, \ldots, Cov(\hat{Y}(r), \widehat{X}_k^r))'$.

 ISBN: 978-84-608-6082-2

Proof.

By squaring both sides in expression (2), taking expectations and neglecting higher order terms we obtain the following approximation

$$V(\widehat{Y}_g^r) = E[\widehat{Y}_g^r - Y]^2 = E\left[\hat{Y}(r) + \sum_{h=1}^{k} G_h'|_{(Y,\mathbf{1})}(u_h - 1) - Y\right]^2. \tag{3}$$

On differentiating (3) and equating to zero, we obtain the optimum values of the parameters as

$$(G_1'|_{(Y,\mathbf{1})}, \ldots, G_k'|_{(Y,\mathbf{1})})' = QD^{-1}b,$$

where $D = (d_{ht})$, $b = (b_1, ..., b_k)'$ and

$$d_{ht} = \frac{Y^2 Cov(\widehat{X}_h, \widehat{X}_t)}{X_h X_t V(\widehat{Y}_\beta)} \quad ; \quad b_h = \frac{Y Cov(\hat{Y}(r), \widehat{X}_h)}{X_h V(\hat{Y}(r))}.$$

On substituting the optimum values into (3) we obtain

$$V_{\min}(\widehat{Y}_g^r) = V(\hat{Y}(r)) - \sigma' \mathbf{\Sigma}^{-1} \sigma = V(\hat{Y}(r))(1 - R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k}),$$

where $R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k}$ is the multiple correlation coefficient. This proofs the Theorem 3.3.

We observe that the multiple correlation coefficient increases with the number of secondary variables and with the number of auxiliary parameters, hence the variance of proposed estimators is a monotone decreasing function of the number of secondary variables.

3.2 The difference estimator

The proposed class of estimator can be used to obtain an optimum difference type estimator. Specifically, let us now consider a choice within the class G of the type

$$G(\hat{Y}(r), u_1, \ldots, u_k) = \hat{Y}(r) + \sum_{h=1}^{k} d_h(u_h - 1)X_h,$$

which yields to the difference estimator

$$\widehat{Y}_{gD} = \hat{Y}(r) + \sum_{h=1}^{k} d_h(X_h - \widehat{X}_h) \tag{4}$$

Theorem 3.4 *The variance of this estimator is given by*

$$V(\widehat{Y}_{gD}) = V(\hat{Y}(r)) + \sum_{h,j=1}^{k} d_h d_j cov\left(\widehat{X}_h, \widehat{X}_j\right) + \sum_{h=1}^{k} d_h cov\left(\hat{Y}(r), \widehat{X}_h\right) \tag{5}$$

The optimum d_h values are: $(d_1, ..., d_k)' = \Sigma^{-1}\sigma$ *and*

$$V_{\min}(\widehat{Y}_{gD}) = V(\hat{Y}(r)) - \sigma'\Sigma^{-1}\sigma.$$

Proof. The optimum d_h values can be obtained by differentiating the above expression of the $V(\widehat{Y}_{gD})$ and equating to zero.

3.3 The ratio estimator

A important case within the general class of estimators $\widehat{Y}_g^r$ is the ratio type estimator, which is given by

$$\widehat{Y}_g^r = \hat{Y}(r) \prod_{h=1}^{k} \left(\frac{X_h}{\widehat{X}_h} \right)^{\alpha_h}. \tag{6}$$

Equation (6) provides a class of estimators with various particular cases. The case $k = 1$ and $\alpha = 1$ coincides with the traditional ratio estimator, whereas the case $\alpha = 0$ gives the naive estimator.

Theorem 3.5 *The optimum values of α_h, with $h = 1, ..., k$, that minimize the variance of $\widehat{Y}_g^r$ are $(\alpha_1, \ldots, \alpha_k)' = A^{-1}A_0$, where $A_0 = (a_{0h})$, $A = (a_{ht})$ and*

$$a_{0h} = \frac{Cov(\widehat{X}_h, \hat{Y}(r))}{YX_h} \quad ; \quad a_{ht} = \frac{Cov(\widehat{X}_h, \widehat{X}_t)}{X_h X_t}.$$

Proof. Following the previous section, we can obtain the optimum values of α_h by differentiating the expression of the $V(\widehat{Y}_g^r)$ and equating to zero.

We have tested the performance of the proposed modified estimator with respect to the criteria: relative bias and efficiency through simulation studies. For this purpose, we consider real and simulated populations in this simulation study. Our simulation studies with several discrete and continuous RR schemes, reveal that there is a decrease in the relative bias and the relative mean square error for all randomized response schemes.

Acknowledgements

This study was partially supported by Ministerio de Educación y Ciencia (grant MTM2015-63609-R and FPU grant program, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía).

References

[1] S.L. Warner, *Randomized response: A survey technique for eliminating evasive answer bias*, Journal of the American Statistical Association **60**(309) (1965) 63–69.

[2] R. Arnab, *Optional randomized response techniques for complex survey designs*, Biom. J. **46**(1) (2004) 114–124.

[3] C.N. Bouza, C.Herrera, P.G. Mitra, *A review of randomized responses procedures: the qualitative variable case*, Investigación Oper. **31**(3) (2010) 240–247.

[4] G.Diana, P.F. Perri, *A calibration-based approach to sensitive data: a simulation study*, Journal of Applied Statistics **39**(1) (2012) 53–65.

[5] A. Chaudhuri, R. Mukherjee, *Randomized response. Theory and techniques.* Statistics: Textbooks and Monographs, 85. New York etc.: Marcel Dekker, Inc. xvi, 162 p., 1988.

[6] S.K. Srivastava, H.S. Jhajj, *A class of estimators of the population mean in survey sampling using auxiliary information*, BiometriKa **68** (1981) 341–343.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A first approach to column updating of NonNegative Matrix Factorization

P. San Juan Sebastián[1], A.M. Vidal[1] and V.M. García-Mollá[1]

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

emails: p.sanjuan@upv.es, avidal@dsic.upv.es, vmgarcia@dsic.upv.es

Abstract

The Non-Negative Matrix Factorization has been widely studied to obtain parts-based representations of data. Several algorithms with different approaches had been presented, but in all of them the matrix to factorize comes from a complete set of data. In this paper, we tackle the problem of updating the factorization, this is, computing a new factorization with data received when the initial factorization has been computed. This may be useful for real-time applications or applications where only a part of the data is available at the beginning of the process. Due to the properties of the updating NNMF problem it can be computed with a parallel approach in order to decrease execution times.

Key words: NNMF, parallel updating NNMF,, multicore approach,

1 Introduction

The Non-Negative Matrix Factorization (NNMF) has become a very important tool in fields such as document clustering, data mining, machine learning, image analysis, audio source separation or bioinformatics [1, 2, 3, 4, 5, 6]. NNMF consists on approximating a matrix $A \in \mathbb{R}^{m \times n}$ by the product of two matrices W and H, with some conditions: all elements of the matrix A are non-negative, and $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ with $k \leq \min(m, n)$ are two lower rank matrices with non-negative elements too, such that $A \approx WH$. The problem can be addressed as the computation of two matrices W, H such that

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

 ISBN: 978-84-608-6082-2

In some applications, all data are not available at the beginning of the factorization process. This might happen, for example, in applications such as real time music content analysis, where new data (the music being produced) may have to be processed when it becomes available [7]. So in order to avoid recomputing the whole factorization each time a new data(or group of data) entry is obtained, we formulate the NNMF updating problem as follows:

Having the matrices W and H as a result from the initial factorization

$$A \approx WH \tag{2}$$

and new data $B \in \mathbb{R}^{m\times r}$, we want to calculate a new NNMF decomposition

$$V = [A \quad B] \approx W_1 H_1 \tag{3}$$

where $W_1 \in \mathbb{R}^{m\times k}$ and $H_1 \in \mathbb{R}^{k\times(n+r)}$.

In this paper we propose and test different algorithms to solve the NNMF updating problem without the need of recalculating the whole factorization. Our motivation is to solve the new factorizations as fast as possible, taking advantage of the previously computed factorization.

2 Proposed Algorithms

The updating problem explained in the previous section has two main approximations: adding one column to the known matrix or adding a block of columns.

2.1 Single column updating

The simplest problem occurs when we want to add only one column to the decomposition. In this case, problem (3) simplifies to $V = [A \quad b] \approx W_1 H_1$ where $b \in \mathbb{R}^m$ and $H_1 \in \mathbb{R}^{k\times(n+1)}$. Our first attempt was to create algorithms to solve the problem directly, then we developed a postprocessing scheme to improve our solutions.

2.1.1 Direct versions

Solving this problem is the simplest way to compute last column x of H_1 which can be the first approach to compute the NNMF updating problem. We reformulate problem (3) as follows:

$$V \approx W_1 H_1 : H_1 = [H \quad x] \wedge W_1 = W \tag{4}$$

where $x \in \mathbb{R}^k$. We developed two algorithms to do so:

1. *NNLS version:* In this algorithm we solved the NonNegative Least Squares problem [8] using our previous factorization matrix W and the new column b. We implemented a MATLAB function to solve the problem.

$$\min_{x \geq 0} \|Wx - b\|_2. \tag{5}$$

2. *LSQ version:* This version solves the unconstrained Least Squares problem (usually computed through a QR decomposition)and then rectifies to 0 all negative values if necessary.

$$\min_{x \in \mathbb{R}^k} \|Wx - b\|_2. \tag{6}$$

Note that this is not the correct solution for the NNMF updating problem, it is only an approximation.We solved the LSQ problem using the MATLAB operation $x = W \backslash b$.

2.1.2 Postprocessed versions

Another option is to compute a few iterations of the complete NNMF algorithm after using the direct versions in order to improve the solution. Furthermore, this method has the advantage of updating W too and not only H.

We used the Multiplicative Algorithm of Lee&Seung [9] using the matrices W and H_1 generated by the direct versions as initialization matrices, particularly a MATLAB implementation that we call MLSA. Using the postprocessing scheme with a randomly generated column x of H_1 was tested too.

Algorithm 1 Postprocesing procedure

1: Compute vector x using one of our direct versions.
2: Execute a few iterations of MLSA with W and $H_1 = [H \quad x]$as initialization values.

2.2 Column block updating

In other applications updating the factorization with more than one data entry is needed. This goal can be achieved with multiple one column updates, or with a block based algorithm which will be more efficient. We created a block based algorithm for our LSQ version solving:

$$\min_{X \in \mathbb{R}^{k \times r}} \|WX - B\|_F. \tag{7}$$

This problem is equivalent to solve r times problem (6), one for each column of B.

 ISBN: 978-84-608-6082-2

However, using a multiple column updating approach has benefits in terms of parallelism. In both versions each single subproblem only depends on the matrix W and one column of B and produces one column of H_1. So each problem can be executed by a single thread in a multicore or GPU environment. In an ideal scenario, this might allow to solve the full updating problem (for many columns) using the same time needed to compute a single column. That will only happen if there are as many cores as columns.

3 Algorithm evaluation

For the algorithm evaluation we used the MLSA to compute the factorizations of matrix A which is the base case. Then we compared 6 algorithms: using MLSA to solve the factorization of V, NNLS version, NNLS version with postprocessing, random version with postprocessing, LSQ version and LSQ version with postprocessing.

3.1 Test1: Fixed base problem size with increase of added columns

In the first test we kept the same problem size (m,n and k) and increased the number of columns to add (r). In the Figure 1 the error results obtained are shown, while Table 1 shows the time spent by the algorithms.

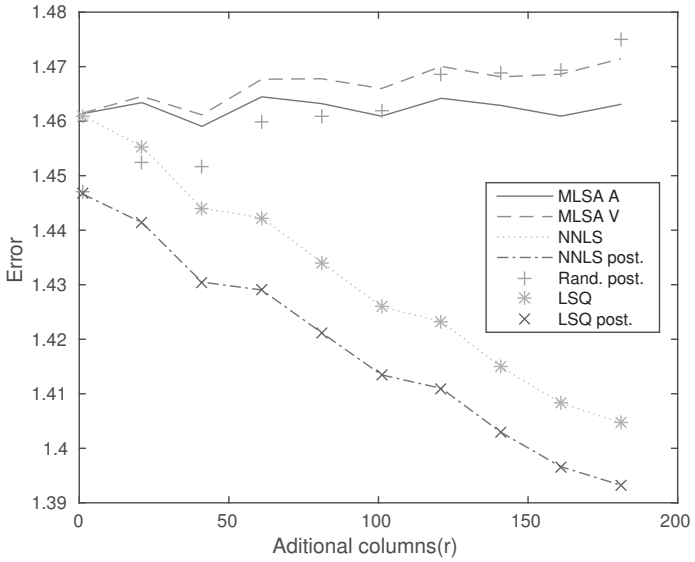


Figure 1: Error results of test 1.

 ISBN: 978-84-608-6082-2

Regarding error, it is easy to see that both postprocessed versions achieve lower error that its simple counterparts. Each version (NNLS and LSQ) has very similar error values, but NNLS version is the one with lower error. The random version has a very good error value for lower values of r, but when it increases the error grows up as initially expected.

Table 1: Execution times of test 1

r **value**	**Algorithm**						
	MLSA A	MLSA V	NNLS	NNLS p.	Rand. p.	LSQ	LSQ p.
$r = 1$	11.197	11.180	3.580	4.465	1.125	0.112	1.233
$r = 21$	10.952	11.316	73.386	76.074	1.199	0.122	1.252
$r = 41$	10.930	11.361	142.934	145.518	1.213	0.118	1.274
$r = 61$	10.952	11.645	212.679	212.576	1.279	0.126	1.239
$r = 81$	11.118	11.589	280.651	280.862	1.207	0.150	1.285
$r = 101$	10.930	12.043	347.219	349.767	1.275	0.133	1.237
$r = 121$	10.849	12.207	417.860	417.579	1.314	0.147	1.263
$r = 141$	10.918	12.483	487.899	487.272	1.317	0.139	1.410
$r = 161$	11.149	12.656	555.071	565.382	1.363	0.137	1.439
$r = 181$	11.132	12.904	636.051	638.566	1.398	0.149	1.379

When taking into account time, NNLS versions time grows too much with the increase of r making it unusable in practice for block column updates. On the other hand, LSQ version has a very low execution time, particularly in its simple version.

Note that as said in previous section, NNLS can be improved to its time with $r = 1$ using a parallel approach. In this experiment, the 600 seconds of the biggest problem tested can be reduced to around 3,5 seconds with 181 cores. With that set-up a speedup of 171,43 can be obtained.

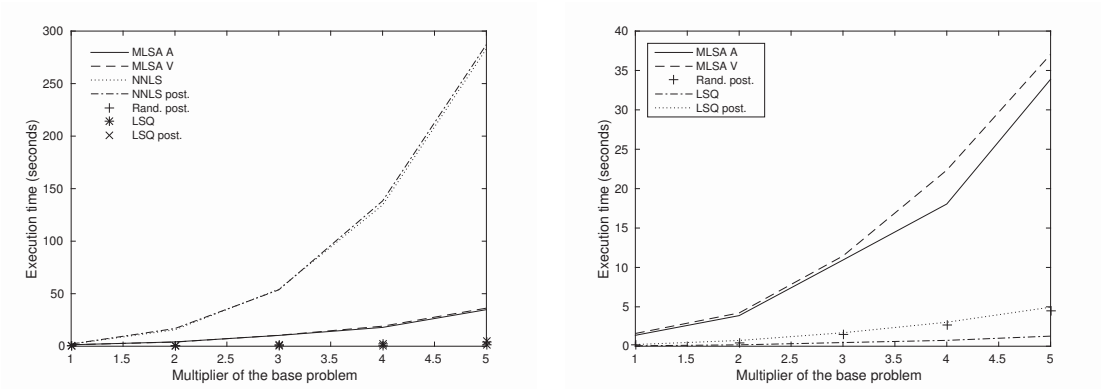
3.2 Test2: Increasing problem size with $r = 1$

In the second test, we kept r fixed and increased the problem size (m, n, k) proportionally. The error measures for each size maintain the same relations of the test 1 shown in Figure 1. The Figure 2(a) illustrates that NNLS and MLSA execution times grow with the problem size, while LSQ keeps almost the same execution time.

3.3 Test3: Increasing problem size with variable r

In the last test, we increased proportionally all problem dimensions(m, n, k, r). As shown in the first test, NNLS time grows too much when we increase r, therefore, NNLS was not executed in this experiment. In Figure 2(b) the execution times of base MLSA, full MLSA, random postprocessed, LSQ and LSQ postprocessed are shown. It can be clearly

seen that as we increase the size, the gap between the full MLSA and the updating methods gets bigger. This means that our updating methods are still better when we tackle bigger problems.



(a) Fixed $r = 1$.

(b) Variable r proportional to problem size.i

Figure 2: Evolution of the execution time whith the increase of problem size.

4 Conclusions and future work

Our experiments showed that the LSQ update algorithm is by far the fastest of all compared algorithms. Despite not being so accurate, the LSQ algorithm produces a very low error in a low execution time. LSQ error is lower than full MLSA error and very close to NNLS error. On the other side, the NNLS update algorithm is too slow to be used in practice, because there is not any gain in execution time against the full MLSA.

In terms of parallelism, both versions are easily to implement for parallel systems. The NNLS version will be hugely benefited by this, because the time for $r = 1$ update is lower than the full MLSA time.

In this paper we used the MLSA as base algorithm, in future works we will test more modern algorithms for the NNMF. Another main goal for us is to develop implementations of these update algorithms for multicore, GPU and Xeon Phi systems.

Acknowledgements

This work has been partially supported by Programa de FPU del Ministerio de Educación, Cultura y Deporte, by Ministerio de Economa y Competitividad from Spain, under the project TEC2015-67387- C4-1-R, and by project PROMETEO FASE II 2014/003 of Generalitat Valenciana.

References

[1] E. Battenberg, A. Freed and D. Wessel, *Advances in the Parallelization of Music and Audio Applications*, Proceedings of the International Computer Music Conference, New York City/Stony Brook, New York, 2010.

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

[3] F.J. Rodriguez-Serrano, J.J.Carabias-Orti, P. Vera-Candeas, T.Virtanen and N. Ruiz-Reyes, *Multiple Instrument Mixtures Source Separation Evaluation Using Instrument-Dependent NMF Models*, LNCS 7191, Springer-Verlag, 2012.

[4] M.W. Berry, M. Browne, A. Langville, V. Pauca and R. Plemmons, *Algorithms and applications for approximate nonnegative matrix factorization*, Comput. Statist. Data Anal., vol. 52, pp. 155-173, 2007.

[5] K. Devajaran, *Nonnegative Matrix Factorization: An Analytical and Interpretative Tool in Computational Biology*, PLoS Comput Biol 4(7): e1000029. doi:10.1371/journal.pcbi.1000029, 2008.

[6] P. San Juan Sebastin, A.M. Vidal, V.M. Garc a-Moll, F.J. Martnez-Zaldvar, J. Ranilla, P. Alonso, M. Alonso-Gonzlez and R. Cortina. *Experiments with the NNMFPACK library: influence of parameter in the NNMF approximation error* Proceedings of the 15th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2015 pag 1023 -1034

[7] Jouni Paulus and Tuomas Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Signal Processing Conference, 2005 13th European*, pages 1–4. IEEE, 2005.

[8] Charles L Lawson and Richard J Hanson. *Solving least squares problems, Chapter 23 Linear Least Squares with linear inequality constraints.* Edited by SIAM, Philadelfia 1995.

[9] D.D. Lee and H.S. Seung, *Algorithms for non-negative matrix factorization*, Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2001.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Accelerating Schur Complement Domain Decompostion Method for Wind Field Calculation

Gemma Sanjuan[1], Tomàs Margalef[1] and Ana Cortés[1]

[1] *High Performance Computing Applications for Science and Engineering research group,*
Computer Architecture & Operating Systems department,
Universitat Autònoma de Barcelona

emails: gemma.sanjuan@caos.uab.es, tomas.margalef@uab.es, ana.cortes@uab.es

Abstract

Wind field calculation is a key issue in many fields, from forest fire propagation prediction to leisure activities planing. However, when the involved terrain map is large the amount of memory required and the execution time become prohibitive to be useful in an operational environment. Domain decomposition methods appear as a promising approach to parallelize and accelerate wind field calculation. Schur complement method has been widely used in finite element problems. In this method, the problem is split into non-overlapping subdomains, and the unknowns in the interiors of the subdomains are eliminated. The remaining Schur complement system on the unknowns associated with subdomain interfaces is solved by the conjugate gradient method. Taking advantage of the particular features of wind field matrices, an approximation has been designed to accelerate the system solution without significant loose of precision in wind field results. The new approach reduces execution time and significantly improve scalability.

Key words: wind field, Schur complement, domain decomposition, Gaussian elimination, diagonal sparse matrix

1 Introduction

Meteorological weather forecast is a complex problem that requires the use of large computing systems to provide accurate predictions meeting time constraints. However, meteorological predictions provide the general trends in the weather situation, but usually, they do not consider local phenomena due to the resolution used when defining the mesh describing the atmosphere (the typical resolution is about some few kilometers). In the particular case

of the wind, the meteorological models provide the wind speed and direction at a $2,5$ - 4 km resolution. However, the meteorological wind is modified by the terrain topography, generating a complete wind field, with different speed and direction at different points of the terrain. So, for several applications, such as forest fire propagation prediction, where the wind effect is critical, it is mandatory to estimate the wind field at high resolution (30 meters or even higher).

So, it is necessary to apply more detailed models to represent and reproduce the variation of the wind components due to the topography of the terrain. Such models, usually, involve solving huge complex systems of equations that take large time to be solved. Therefore, it is necessary to apply parallelization and high performance computing techniques to reduce execution time and make the use of such systems feasible in real-time.

Domain decomposition has been applied to many problems considering finite elements and has been parallelized in many different ways [11][6], but in some cases, the particular conditions of the problem being considered can be exploited to simplify and accelerate the solving process. This is the particular case of wind field, where the matrices involved present some particular features that can be exploited. The same features have also been found in geophysical problems [3]. So, the methodology developed can be applied to a full set of problems involving the same kind of matrices.

Section 2 presents a wind field simulator, called WindNinja, and analyses the method applied to build the mesh, the system of equations and the way of solving it. Section 3 introduces the application of the Schur Complement Domain Decomposition Method to parallelize the solution. This method parallelizes quite well, but the scalability has several limitations that can be overcome by taking advantage of particular features of the involved matrices. So, Section 4 presents the approximation proposed in this work to improve execution time and compares the results with those obtained by the original solution. The approach is combined with a previously developed sparse matrix-vector multiplication method to accelerate the Preconditioned Conjugate Gradient used to solve the system of equations. 5 summarizes some results, comparing execution time and quality of solution, and, finally, Section 6 summarizes the main contributions of this work.

2 WindNinja - Wind Field Simulator

WindNinja [5] is a wind field simulator that takes the elevation map and the meteorological wind speed (ws) and direction (wd) and determines the wind values at each cell of the terrain. It is based on the equations that describe air flow variation in the atmosphere. Specifically, it is a mass-consistent model initialized by boundary conditions. The function to minimize is constructed using the square of the difference between the adjusted and observed values as shown in Equation 1,

$$E(u,v,w) = \int_{\Omega} [\alpha_1^2(u-u_0)^2 + \alpha_1^2(v-v_0)^2 + \alpha_2^2(w-w_0)^2]d\Omega \quad (1)$$

where *u, v, w* are the velocity components in the x (positive to East), y (positive to North), and z (positive upward) directions, respectively; u_0, v_0, w_0 are initial values of velocity, and α_i is the Gauss precision moduli that can be used to control the relative amount of change induced by the model to the horizontal and vertical directions.

The minimization of Equation 1 is subject to the strong constraint of conservation of mass that can be expressed as shown in Equation 2.

$$H(u,v,w) = (\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}) = 0 \quad (2)$$

This constraint can be expressed applying Lagrange multiplier theory. So, the function can be expressed as shown in Equation 3,

$$F(u,v,w,\lambda) = \int_{\Omega} [\alpha_1^2(u-u_0)^2 + \alpha_1^2(v-v_0)^2 + \alpha_2^2(w-w_0)^2 + \lambda(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z})]d\Omega \quad (3)$$

where $\lambda(x,y,z)$ is the Lagrange multiplier.

This equation can be expressed as a linear system in the form:

$$Ax = b, \quad (4)$$

where A is the matrix formed from the mesh of points, b are the Dirichlet boundary conditions and x are the unknowns of the system. In the linear system $Ax = b$, the A matrix is a sparse matrix which is stored in CRS (Compressed Row Storage) format. The matrix A has a low density and a diagonal pattern with elements in the main diagonal and just 26 subdiagonals (13 considering that it is symmetric). This system cannot be solved by direct methods [4] and iterative methods must be applied [8]. WindNinja applies the preconditioned conjugate gradient (PCG) [7] solver to solve the system of equations. The PCG is an iterative method that uses a matrix M as a preconditioner and iteratively approaches the solution. Matrices A and M are constant during the whole iterative process.

The main calculation involved in this process is the PCG solver applied to solve the system of equations. The x vector, which is the solution to be reached, lies at the intersection point of all the hyperplanes created by the quadratic form of each equation of the equations system. To reach this value, x is initialized at x_0 and, at each iteration, it is modified to approach the real solution. The way x is modified is explained by following Algorithm 1.

Depending on the preconditioner, the solver takes more or less iterations to converge and the cost of each iteration is larger or smaller. However, the main problem of is that the

Starting from x_0
Calculate $g_0 = Ax_0 - b$,
which is the difference between the initial value and the real value.
Considering that M is the preconditioner,
evaluate $q_0 = M^{-1}g_0$ and
set the initial value of p as $p_0 = -q_0$

For k=1, ..., n:

$$\alpha_k = \frac{(g_k, q_k)}{(p_k, Ap_k)}$$
$$x_{k+1} = x_k + \alpha_k p_k$$
$$g_{k+1} = g_k - \alpha_k A p_k$$
$$q_{k+1} = M^{-1} g_{k+1}$$
$$\beta_k = \frac{(g_{k+1}, q_{k+1})}{(g_k, q_k)}$$
$$p_{k+1} = q_{k+1} + \beta_k p_k$$

Algorithm 1: Preconditioned Conjugate Gradient (PCG)

matrices involved for real maps are extremely large, and the matrix operations takes very long time. So, domain decomposition methods has been applied to reduce the size of the problem and accelerate the resolution of the system.

3 Schur Complement Domain Decomposition Method

The goal consists of solving a linear system: $Ax = b$, where A is a symmetric and positive definite sparse matrix. The linear system is obtained from the partial differential equation representing the mesh. The problem to be solved, applying the Schur Complement Domain Decomposition method, is partitioned into a certain number of subdomains without overlapping, plus one interface subdomain s. The interface s is defined so that for any pair of subdomains p_i and p_j, no node of p_i is directly coupled with any node of p_j, but only with its own nodes and with s. Figure 1 shows an example of the domain decomposition. In Figure 1.a) the original mesh, showing the interactions among the nodes is introduced. Figure 1.b) shows the domain decomposition in 3 domains (0, 1 and 2) and 2 interfaces (S_1 and S_2). In this case, the interface numbered as S_1 is the interface between subdomains 0 and 1, and the interface numbered as S_2 is the interface between subdomains 1 and 2. This interactions are distributed in parts of matrix A as indicated in Figure 1.c). In this figure, $A_{i,i}^{(II)}$ indicates the submatrix that contains the internal interacions among the nodes

of subdomain i (thick lines); $A_{i,S_j}^{(IS)}$ indicates the submatrix that contains the interactions among the nodes in subdomain i and the nodes in the interface S_j (dotted lines); and $A_{S_j}^{(SS)}$ indicates the submatrix that contains the interactions among the nodes in the interface S_j (dashed lines).

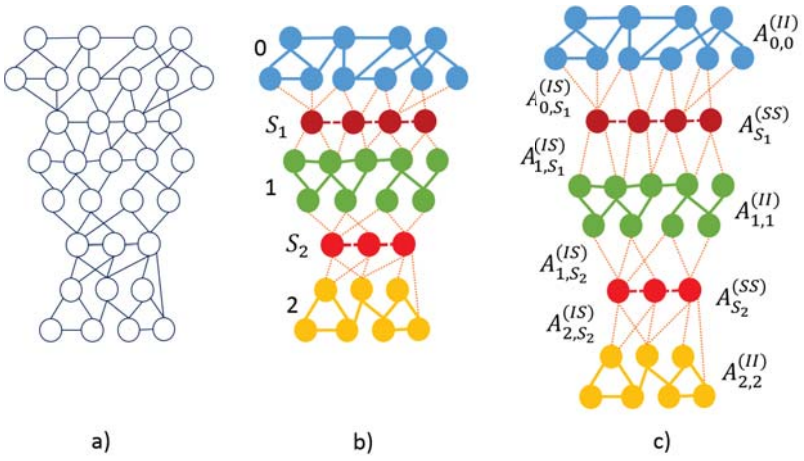


Figure 1: Mesh partitioning in the Schur Complement Domain Decomposition method

This matrix structure is represented in Figure 2.

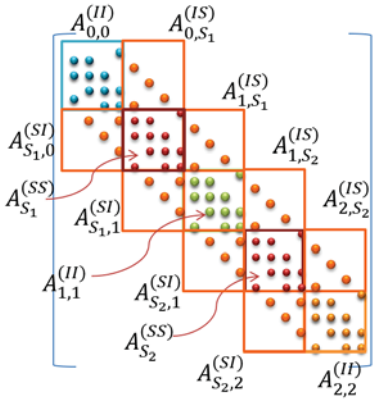


Figure 2: Matrix A considering Schur Complement Domain Decomposition method

This matrix can be reorganised, so that the internal subdomain interactions are grouped and the internal interface interactions are also grouped. With this reorganisation, the matrix can be represented as shown in Figure 3a. In this matrix, all the internal interfaces

interactions can be joined in a single square submatrix and all the interactions between one subdomain and the interface nodes can also be grouped in a single submatrix. The final structure of matrix A is shown in Figure 3b.

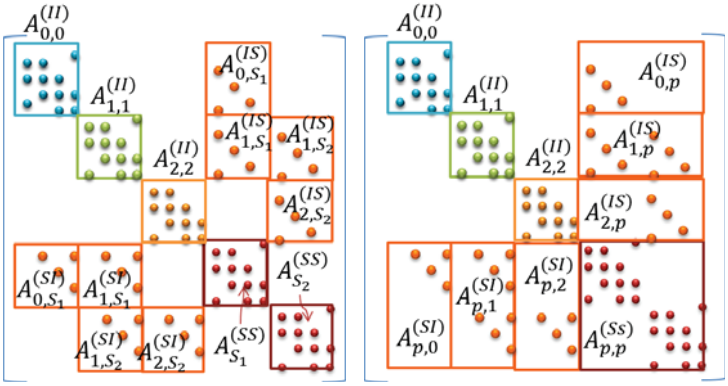


Figure 3: Reorganization of Matrix A

The final matrix A structure can be described as follows:

1. A set of square sparse submatrices $(A_{0,0}^{(II)} \dots A_{p-1,p-1}^{(II)})$, each one representing the interactions among the nodes in a subdomain. The dimension of each one of these submatrices depends on the number of nodes in that particular subdomain. These matrices are organised along the diagonal of the system.

2. A square sparse submatrix $(A_{p,p}^{(SS)})$ representing the interactions among the nodes in the interface. This submatrix is located at the bottom of the diagonal of the system.

3. A set of rectangular submatrices located in the last column $(A_{0,p}^{(IS)} \dots A_{p-1,p}^{(IS)})$ of the system (and the corresponding transposed ones), representing the interactions among the nodes in one particular subdomain and the nodes in the interface subdomain.

4. The rest of the elements of the matrix system are zero, because there are no interaction among the nodes of one particular subdomain and the nodes of any other subdomain.

Considering this organisation, and applying the Gaussian elimination, the system becomes the following one:

$$\begin{pmatrix} A_{0,0}^{(II)} & 0 & \cdots & A_{0,p}^{(IS)} \\ 0 & A_{1,1}^{(II)} & \cdots & A_{1,p}^{(IS)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S \end{pmatrix} \begin{pmatrix} x_0^{(I)} \\ x_1^{(I)} \\ \vdots \\ x_p^S \end{pmatrix} = \begin{pmatrix} b_0^{(I)} \\ b_1^{(I)} \\ \vdots \\ s_S \end{pmatrix} \tag{5}$$

where S and s_S are, respectively:

$$S = A_{p,p}^{(SS)} - \sum_{i=0}^{p-1} A_{p,i}^{(SI)} A_{i,i}^{(II)-1} A_{i,p}^{(IS)} \tag{6}$$

$$s_S = b_p^{(S)} - \sum_{i=0}^{p-1} A_{p,i}^{(SI)} A_{i,i}^{(II)-1} b_i^{(I)} \tag{7}$$

To solve the system, the first step is to solve the bottom subsystem:

$$S x_p^{(S)} = s_S \tag{8}$$

This system is much smaller than the original system and can be solved by applying the preconditioned conjugate gradient solver much faster than the original system $Ax = b$. To solve the system by applying a PCG solver, it is necessary to evaluate $w = Sp_k$ (Ap_k in Algorithm 1). This matrix-vector multiplication can be expressed as:

$$w = A_{p,p}^{(SS)} p_k - \sum_{i=0}^{p-1} A_{p,i}^{(SI)} A_{i,i}^{(II)-1} A_{i,p}^{(IS)} p_k \tag{9}$$

So, this operation is decomposed into a set of terms of a sum that can be calculated in parallel. The resulting vectors of each multiplication can be added, and the result is used in the following step of the PCG algorithm to obtain the new value of p_{k+1}. This process is iterated until the system is solved. Once $x_p^{(S)}$ has been obtained, all the $x_i^{(I)}$ can be solved simultaneously from the original equation (5).

$$x_i^{(I)} = A_{i,i}^{(II)-1} (b_i^{(I)} - A_{s,i}^{(IS)} x_p^{(S)}) \tag{10}$$

The Shur method presents several advantages and possibilities of exploiting parallelism. So, the method has been implemented as a MPI Master-Worker application. First, the Master process partitions the mesh and distributes the submatrices corresponding each subdomain to one different worker. So, one worker receives the matrix describing the internal interactions of that subdomain and the interactions among that subdomain and the intereface. Each worker evaluates one of the terms of the sum of Equation 6 calculating the inverse matrix involved. Then, the worker multiplies the result by vector p_k and sends

the resulting vector to the Master process. The Master process carries out one of the iterations of the PCG algorithm and sends the new p_{k+1} to the workers. The workers evaluate the new matrix-vector multiplication and send back the result to the Master. This process is repeated to solve $x_p^{(S)}$. Once it has been solved the result is sent to the workers that calculate the remaining $x_i^{(I)}$ terms of the solution.

4 Accelerating Schur Complement Domain Decomposition Method

In this process, there are two operations that take most of the time.

1. Inverse matrix calculation: To calculate each one of the corresponding term of Equation 9

$$A_{p,i}^{(SI)} A_{i,i}^{(II)-1} A_{i,p}^{(IS)} p_k, \tag{11}$$

the first step is to calculate the inverse matrix of $A_{i,i}^{(II)}$, and this matrix inversion is a time consuming operation. Actually, this operation and the complete calculation of the matrix product $A_{p,i}^{(SI)} A_{i,i}^{(II)-1} A_{i,p}^{(IS)}$ must be only carried out once at the beginning of the process.

Several libraries have been tried to calculate the inverse matrix. The most successful one was PARDISO [1], that provided the best execution times for WindNinja submatrices. The PARDISO inversion algorithm first constructs the LDL^T factorization of matrix A, where L is the Cholesky factor, and D is a block diagonal matrix. The inversion method used in PARDISO for symmetric matrix does not provide the exact inverse matrix, but it makes an accurate approximation.

Another approach could be to apply a LU factorization. In this way, each term shown in Equation 11 can be calculated without calculating the inverse matrix itself.

Moreover, as mentioned above, WindNinja matrices have very specific features that can be exploited to accelerate this step. WindNinja matrices have the main diagonal and 26 subdiagonals, that can be at different positions of the matrix. The main diagonal has really significant values, and the first subdiagonal, also has significant values, but the other 12 subdiagonals have much less significant values. The farther the subdiagonal, the lower significance. So, it would be worthy to study the effect of discarding the less significant subdiagonals from each matrix $A_{i,i}^{(II)}$ in the final wind field. This approach can reduce the execution time of the LU factorization. So, an study has been carried out to determine the execution time and wind field difference (measured as root mean square error RMSE in the wind speed in m/s),

when discarding different number of diagonals. These measures for a 800x800 cells map are shown in Table 1.

Disc. Subdia.	Exec. Time (s)	RMSE (m/s)
2	2461.00	1.24
4	700.89	0.93
6	310.26	0.81
8	200.22	0.67
10	140.25	0.46
12	90.20	0.35

Table 1: WindNinja Execution time and RMSE when discarding different number of subdiagonals

In this table, *Disc. Subdia.* means the number of discarded subdiagonals in the upper triangular and in the lower triangular. So, a value of 12 means that 24 subdaigonals are discarded and the resulting matrix only has the main diagonal, one upper subdiagonal and one lower subdiagonal. It can be observed that the execution time is significantly reduced (from 2461.00 seconds to just 90.20 seconds) and the error introduced in the wind field is not very significant.

2. Sparse Matrix-Vector multiplication:

   The calculation to obtain $A_{p,i}^{(SI)} A_{i,i}^{(II)-1} A_{i,p}^{(IS)}$ is carried just once for each $A_{i,i}^{(II)}$ matrix. Once this matrix multiplication has been carried out, it is necessary to carry out a sparse matrix-vector multiplication to obtain $A_{p,i}^{(SI)} A_{i,i}^{(II)-1} A_{i,p}^{(IS)} p_k$. And this sparse matrix-vector multiplication ($SpMV$) must be repeated for iteration of the PCG algorithm. So, the performance of this operation is critical to improve the performance of the PCG, and the performance of WindNinja itself.

   The $SpMV$ operation has been widely studied and many implementations considering different approaches can be found in the literature [2] [10]. But, most of this approaches do not work properly with WindNinja matrices due to the extremely low density of the matrices involved (matrix density is usually below 0.001. So, in a previous work, the authors proposed an storage format, called $VDSpM$) and a sparse matrix-vector multiplication method ($VDSpMV$) based on that storage format, that reduce the $SpMV$ with such low density [9] exploiting parallelism at node level. So, This method has been applied to accelerate $SpMV$ and improve WindNinja excution time. This method introduces an OpenMP parallelization that allows to exploit the cores available in the current computing nodes, with a significant SpeedUp up to $4-8$ cores.

5 Experimental results

A wide set of experiments has been carried out, considering different map size. In this experiments, three implementations of WindNinja have been executed. The original sequential WindNinja, the Schur complement domain decomposition method using different number of subdomains and PARDISO library and the Schur complement domain decomposition method using subdiagonal discard, LU factorization, Gaussian elimination and $VDSpMV$ multiplication method. Table 2 presents a summary of these experiments. In this table, $Map\ Size$ is the size of the map in number of rows and columns, $Subdom.$ is the number of subdomains used to solve the system, $WindNinja$ is the execution time of WindNinja on a single core, $Time\ P$ is the execution time applying domain decoposition, and using PARDISO to calculate the inverse matrix, $Time\ D$ is the execution time of WindNinja considering subdiagonal discarding, $RMSE\ P$ is the root mean square error in the wind speed obtained when applying PARDISO and $RMSE\ D$ is the root mean square error in the wind speed obtained when applying subdiagonal discarding.

Map size	Subdom.	WindNinja (s)	Time P (s)	Time D (s)	RMSE P (m/s)	RMSE D (m/s)
200x200	1	40	-	-	-	-
200x200	2	-	387	19	1.418	1.790
200x200	4	-	197	17	0.907	1.145
200x200	8	-	77	13	0.364	0.321
200x200	10	-	42	18	0.327	0.186
200x200	12	-	57	22	0.260	0.176
200x200	16	-	57	30	0.022	0.010
400x400	1	102	-	-	-	-
400x400	2	-	510	43	1.702	1.969
400x400	4	-	232	38	0.817	1.031
400x400	8	-	92	29	0.328	0.289
400x400	10	-	43	41	0.321	0.182
400x400	12	-	56	50	0.234	0.123
400x400	16	-	70	68	0.009	0.098
800x800	1	569	-	-	-	-
800x800	2	-	6209	125	1.532	1.772
800x800	4	-	1306	105	0.985	1.134
800x800	8	-	186	60	0.358	0.315
800x800	10	-	88	62	0.294	0.167
800x800	12	-	69	68	0.256	0.134

Continue in next page

Map size	Subdom.	WindNinja (s)	Time P (s)	Time D (s)	RMSE P (m/s)	RMSE D (m/s)
800x800	16	-	114	84	0.026	0.011
1000x1000	1	1459	-	-	-	-
1000x1000	2	-	408	341	1.392	1.757
1000x1000	4	-	3901	327	0.891	1.124
1000x100	8	-	206	298	0.437	0.353
1000x1000	10	-	103	203	0.353	0.184
1000x1000	12	-	111	213	0.313	0.150
1000x1000	16	-	538	339	0.023	0.010

Table 2: WindNinja Execution time and RMSE considering different implementations

These results show that Schur complement domain decomposition method reaches a successful scalability up-to 10 subdomains. Using PARDISO to calculate the inverse matrix requires too much time when the matrices involved are large. When the number of subdomains is increased, the size of the matrices to be inverted is reduced and the time is reduced. However, when the number of subdomains increases the size of the interface increases and solving the interface subsystem by the PCG increases its execution time, penalizing the overall execution time. The method proposed in this work, with subdiagonals discarding, reduces execution time significantly, without introducing a large error in the final wind field.

6 Conclusions

A Schur complement domain decomposition method has been applied to calculate wind field at high resolution for large terrain maps. The method provides a successful scalability up-to 10 subdomains, because, the interface part increases with the number of subdomains. The method introduces natural parallelism that can be exploited at cluster and node level. The final approach is a hybrid application that first solve the interface subdomain reducing the execution time by using a subdiagonal discarding methodology, and afterwards solve the remaining subdomains in parallel.

Acknowledgements

This work has been supported by Ministerio de Economía y Competitividad under contract TIN-2014-53234-C2-1-R and partly supported by the European Union FEDER (CAPAP-H5 network TIN2014-53522-REDT).

References

[1] PARDISO 5.0.0 Solver Project. http://http://www.pardiso-project.org/, 2008. [Online; accessed 12-May-2016].

[2] Kadir Akbudak, Enver Kayaaslan, and Cevdet Aykanat. Hypergraph partitioning based models and methods for exploiting cache locality in sparse matrix-vector multiplication. *SIAM Journal on Scientific Computing*, 35(3):C237–C262, 2013.

[3] S.R. Bernabeu, V. Puzyrev, M. Hanzich, and S. Fernandez. Efficient sparse matrix-vector multiplication for geophysical electromagnetic codes on xeon phi coprocessors. pages 61–65, 2015. cited By 0.

[4] T. Davis. *Direct Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2006.

[5] J.M. Forthofer, K. Shannon, and B. W. Butler. Initialization of high resolution surface wind simulations using nws gridded data. In *Proceedings of 3rd Fire Behavior and Fuels Conference; 25-29 October*, 2010.

[6] Ruipeng Li, Yuanzhe Xi, and Yousef Saad. Schur complement based domain decomposition preconditioners with low-rank corrections. *CoRR*, abs/1505.04340, 2015.

[7] Jorge Nocedal and Stephen J Wright. *Conjugate gradient methods*. Springer, 2006.

[8] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003.

[9] Gemma Sanjuan, Carles Tena, Tomàs Margalef, and Ana Cortés. Applying vectorization of diagonal sparse matrix to accelerate wind field calculation. *The Journal of Supercomputing*, pages 1–19, 2016.

[10] Samuel Williams, Leonid Oliker, Richard Vuduc, John Shalf, Katherine Yelick, and James Demmel. Optimization of sparse matrix-vector multiplication on emerging multicore platforms. In *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing*, SC '07, pages 38:1–38:12, New York, NY, USA, 2007. ACM.

[11] G. Yagawa, N. Soneda, and S. Yoshimura. A large scale finite element analysis using domain decomposition method on a parallel computer. *Computers Structures*, 38(5):615 – 625, 1991.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Certificate Graph Based Authentication for Communications in Emergency Situations

Iván Santos-González, Pino Caballero-Gil, Jezabel Molina-Gil and Alexandra Rivero-García

Department of Computer Engineering and Systems, Universidad de La Laguna

emails: jsantosg@ull.edu.es, pcaballe@ull.edu.es, jmmolina@ull.edu.es, ariverog@ull.edu.es

Abstract

This work presents a new communication system for emergency situations or natural disasters, where the use of network infrastructures is not possible because they are collapsed or unavailable. This proposal uses different wireless technologies such as Bluetooth Low Energy, Wi-Fi Direct and LTE Direct in order to enable Peer-to-Peer communications between users through their smartphones. The general procedure of the proposed system is based on the use of public chat rooms, with open access for all registered and authenticated users, and private chat rooms that can be used by different groups of people. These permissions are established using a distributed authentication scheme based on public keys and certificate graphs. The users cooperate in the message forwarding to other users. Moreover, this system includes a decision protocol to select the most appropriate communication technology in every moment.

Key words: Natural Disaster, Communication, Location, LTE Direct, Wi-Fi Direct, Bluetooth Low Energy

1 Introduction

Nowadays, natural disaster and emergency situations continue being one of the fields where communications can get interrupted because it is impossible to be completely ready for them. The most common communication problem in these situations corresponds to when the network infrastructure becomes unavailable or collapsed. This communication interruption is an important problem for the people who are in in emergency situations because they could be isolated, but even more for the organizations responsible for the security and

health of the people that are suffering the consequences derived from the disaster. In these unpredictable situations, it is unrealistic to assume that everybody will carry some special material or gadget. However, a gadget that everybody usually carries most of the time is the smartphone. The use of smartphones is nowadays widespread and extended so that most people consider them essential in their daily lives because smartphones are now used for practically everything from calling and texting, to email, video viewing, Internet access, social media, geolocation, photography, etc. Due to this, in this paper a new decentralized communication system based on smartphones and formed by multiple wireless technologies is proposed for its use in emergency situations where the network infrastructures can not be used.

The proposal we present consists on the alternative use of different technologies, such as Bluetooth Low Energy (BLE), Wi-Fi Direct or LTE Direct through the users smartphones, when in the case of emergencies the network infrastructure is not available. The general procedure of the system is base on the use of public chat rooms where we have access when are authenticated and registered using asymmetric cryptography where distributed public key certification is faced through certificate graphs. Moreover, it is predicted that the different authorities who have a indispensable role in cases of emergency, can use the system to access to private chat rooms thanks to the decryption using the secret key shared in groups.

The present work is structured as follows. Section II describes a brief state of the art. The proposed system is defined in Section III. Section IV introduces the distributed authentication based on certificate graphs used in the proposal to improve the security. Finally, some conclusions and open issues close the proposal in Section V.

2 Preliminaries

During the last years, different proposals have been presented in the field of communications in emergency situations. A lot of them have focused their systems in the use of the emergency call, included in all mobile devices some years ago. This mode allows the users to make an emergency call even when they have not coverage to make a phone call to any of their contacts. To make this system works, The European Union imposed in 1991 to its members states the use of the 112 number to make emergency calls. Moreover it imposed to mobile operators and to mobile phone manufactures the adaptation of all their products in order to, independently of the mobile operator to which the user belongs, let the users use the network infrastructure of any other operator to make emergency calls. In this fact is based Alpify [1], that allows the users to, using the emergency call system in cases where there are not Internet connection, notify through SMS that includes the georeferenced coordinates of the position where the injured person is to both, the emergency services and a person previously selected in the application to be notified in this kind of situations. A

problem of this proposal that derives from the use of the emergency call system, is the necessity of any network infrastructure to the correct work of the application. It implies to have coverage of at least of an operator. In the proposed situation, the initial hypothesis is that there is not coverage of any mobile operator because the network infrastructures are unusable, collapsed, or do not exist because the users are in remote or in difficult access places. In these cases, the use of any system based on communications using network infrastructures is not a feasible solution to the proposed problem. Moreover, the use of systems that need previous agreements between company that develops the system and the different emergency services of every state of each country, implies difficulty to use them in all places around the world. On the other hand, centralized systems like Alpify only notifies the emergency services and the contact person previously selected, and depending on where they are, they can delay to arrive to the position of the person in an emergency situation. The fact of not letting people who are near help in a faster way, suppose an important disadvantage of this kind of systems with other ones.

Other systems, like the one in [2], propose the use of a wireless technology, Wi-Fi Direct, but in the case where there is not Internet connection we can notify the people that are in the range of the Wi-Fi Direct technology. The behaviour of this system consists on sending an emergency message that includes the geolocated position of the person, to other people in the range of Wi-Fi Direct. When the emergency message is received, if the user has Internet connection he/she notifies the emergency automatically to the emergency services. If the user has not Internet connection, the message is re-sent through a multi-hop system to every near user unless the message was received before. This system involves an improvement compared to other systems than only use the emergency call, but it has problems when the number of users is not enough or they do not have the Wi-Fi Direct technology in their smartphones.

Other systems try to solve this problem through the use of other devices like laptops that are carried by the emergency services, as happen in [3]. In this proposal the intention is that the first units of the emergency services that arrive to the place deploy a MANET based on Wi-Fi through the laptops creating a multi-hop ad-hoc network. After the network deployment, a group communication system based on Peer-to-Peer, which admits communications such as VoIP, Push-to-Talk, instant messaging, social networks, etc. is proposed. Moreover, they propose the use of an information system to rescue people in earthquakes that allows them to manage all the logistics needed in this kind of catastrophes. The main advantages that this kind of solutions offers are that the network is managed by specialized personnel that we suppose have a previous learning, preventing the access of other users to the network because the use is restricted only to emergency services. On the other hand, the people who have suffered the catastrophe and that could be isolated, continue being isolated, both during the time that the emergency services takes to arrive to the catastrophe place, that can be of vital importance, and once that the emergency service are acting. This

fact assuming the privation of help by part of other close people that although have suffered the catastrophe, are in condition to help if it is necessary.

3 Proposed System

Communications in emergency situations and/or natural disasters have been usually a problem because the communication systems have not been thought to operate without network infrastructures. For this reason this work proposes a new communication system for situations like those ones, where the network infrastructures can become unavailable. This system uses different wireless technologies (BLE, Wi-Fi Direct and LET Direct) to, independently of the used technology in every moment, can establish a communication with users that are in the range of these technologies. The proposal works in the same way, independently of the communication technology of the three available in the system that is used in every moment.

The new system offers the users the possibility of communicating directly with other users or using the different chat rooms available for the global group communication. The different chat rooms are available for the users to exchange messages and to organize the help tasks of the emergency situations, in a way that the done tasks are stored for the later revision of other users. Public chat rooms can be used by everyone, but private ones can be only used by people or groups of people with some permissions. A good example of the procedure of this system is the fact that some chat rooms are only available for emergency services like the fire-fighters, the police, health personnel, etc. On this way, only the people who belong to these groups and who have the necessary permissions, can access, read and answer messages of these private chat rooms. Figure 1 shows the general use of our system in a emergency scenario.

In order to make possible that the system can work, all users store in their smartphones a ciphered copy of all chat rooms, including private chat rooms, in their states at the moment of the last connection. Thus, when any device discovers nearby devices it establishes a communication, like synchronizes its local copies to the most updated state and if the user has Internet connection, the local copies gets synchronized with the master copy stored in the central server.

When the users download the application, they only have permission to use the public chat room and to do direct communications with other users. If a user wants to get access to private chat rooms, in case of having Internet connection, fact that happens when he/she downloads the application from the market, he/she has to send a request to the server and send the membership group credentials, and the administrator decides whether to grant access. Once the user is registered in the system, he/she receives part of the central server a pair of public/private keys. In cases where the administrator approves the membership of this user in some of private groups, he/she receives the secret key used to encrypt the chat

Figure 1: Possible Situation

rooms that corresponds to these groups. If the user has no Internet connection, it can send a request to another user with the needed permissions or with administrator permissions for this user to authenticate his/her as system member and certifies him/her as member of a group or not. This is done in this way because in emergency situations where usually there is no Internet connection, the application could be shared between users, and by default all users only have public permission, fact that involves that the users that receive the application only can use public chat rooms and direct communication. Thus, in this way, authenticated users could authenticate others using an interactive strong identification scheme based on a challenge-response scheme with digital signature [4], without the necessity of any of them having Internet connection. Once a device has Internet connection, the system communicates with the server to inform about whether it has been authenticated or it has authenticated by/to other user and then the server decides whether to revoke this authentication or not. The user certification scheme is described widely in Section 4.

Every user is identified by his/her MAC or IMEI and his/her name in the application. Moreover, every chat room belongs to a group of users and nobody has access to a chat room if he/she does not belong to this group. Every group has a secret key, and the users can belong to multiple groups. By default, all users belong to the group with access, read and write permissions for the public chat rooms. When a user is authenticated, he/she receives the secret keys of all existent chat rooms that he/she has permissions. On this way, the users can read the messages published in the chat rooms, including the messages published before. In the synchronization cases, the users cooperate with the system forwarding the received messages and synchronizing their local copies and messages with the copies of other users and the central server, in case of Internet connection, even though the users have not the permissions of a determined group. In Figure 2, we can see the proposed key sharing scheme where we present the case of 4 users A, B, C and D that have access to the public chat room. The arrows represent a communication, and for this communication the users have to be connected. In the case of users A and B, both have Internet connection, and in the moment of the register and authentication with the system central server, they receive from it their keys, and the keys corresponding to the chat rooms that they have access, which in the image are 1 chat room for user A and 2 for user B. Users C and D have not Internet connection and therefore they have to register and authenticate with user B, and receive from him/her their keys and the keys that correspond with the chat rooms that they have access, that are one and two, respectively.

In order to choose which technology must be used in each communication, an automatic decision protocol has been designed. This protocol allows users not to have to select the technology because it automatically chooses the most appropriate one for the communication in each moment. The developed protocol is based on different variables such as the number of peers found in the last successful attempts (established during a period defined by the user, which by default is 30 min), a parameter that depends on how recent are the last successful attempts, to prioritize the most recent attempts, and the consumption weight of each technology.

In the first use of the protocol, since there have not been any previous attempt, it is checked what is the available technology that has the highest number of possible users. In the next attempts, the technology that produces the best results according to the aforementioned parameters is prioritized. Besides, after a number (defined by the user, and established by default as 10) of continued communications using the same technology, the system forces the attempt to use a different technology. On this way, we avoid that a technology that can be promising but in previous attempts have not been used, would not be marginalized. In particular, to choose the technology to be used in each moment, the following formula to compute the weight of each technology, W, is used, where i represents each one of the last 10 attempts with that technology, the Boolean variable b indicates whether the technology was successfully used (1) or not (0) in the attempt i, N_i is the number of

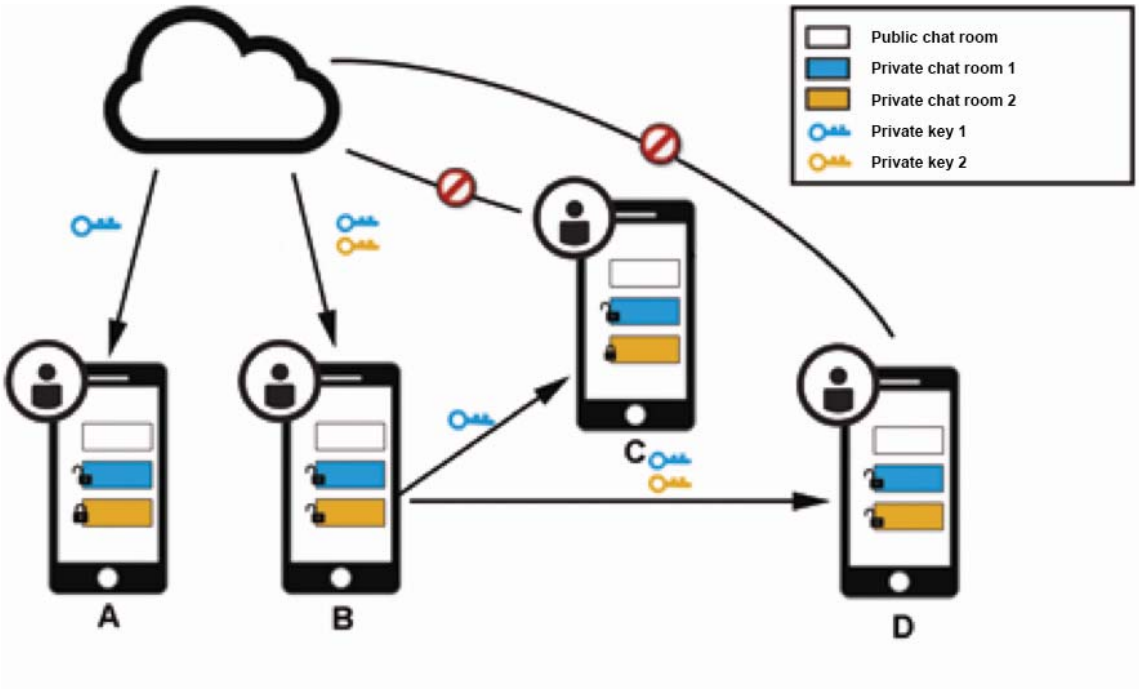


Figure 2: Key Sharing Scheme

peers discovered in the attempt i with that technology, and W_c represents the consumption weight of that technology.

$$W = \frac{\sum_{i=1}^{10} i \cdot b \cdot N_i}{W_c}$$

The consumption weight W_c has been computed for every technology after a consumption study based on a benchmark application for the BLE and Wi-Fi Direct technologies and assuming that the consumption of LTE Direct is similar that the one of BLE as it is describe in [5]. The decision function to select a technology is defined after applying the above formula on each one of the three considered technologies so that the technology that produces the highest result of the formula is the one to be used for the communication.

4 Certificate Verification

In cases of no connection with the central server, the used distributed authentication procedure is based on the concept of certificate graph [6]. A certificate graph is a directed graph

that represents public keys and certificates. In particular, every vertex u of this graph, represents a public key associated to a node, and every edge (u, v) represents a certificate associated with the public key of v, signed with the private key of u. In this graph, a sequence of certificates $P_{uv} = \{(u, u_0), (u_0, u_1), ..., (u_m, v)\}$, where all vertices are different, is named chain of certificates of u to v.

A certificate graph has the next properties:

- The node that emits the certificate is known as issuer, while the node that is certificate is known as subject.
- If a node u knows the public key of another node v, then u can issue a certificate that identifies the public key of v, KU_v.
- If a node w knows the public key of another node v, KU_v, then it can decrypt the certificate from v to u and so obtain the public key of u.
- The length of the chain is the number of certificates in the chain.
- A chain from u to v, is the shortest in the graph if there are not any other shorter chain from u to v in the graph.

In Figure 3 we can see the certificates stored in each node of a simple certificate graph.

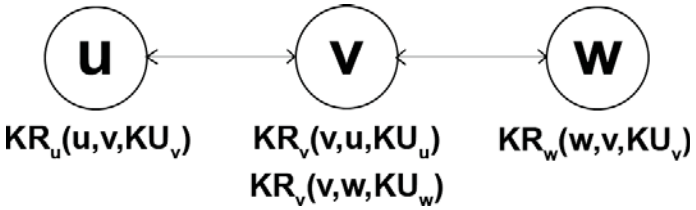


Figure 3: Certificate Graph Example

Each device has an identification associated with its MAC or IMEI (n) that grants access to the chat rooms if it is updated. Moreover, each node has a pair of public and private keys (KU_n, KS_n) created in a centralized way by itself as well as its own certificate. This certificate is emitted by the CA and confirms that the public key that has this certificate, is associated to the node that carry it because it is the only one that has the correspondent private key.

$$Cert_n = (KR_{CA}(KU_n), KU_{CA})$$

The CA is the responsible of maintaining the association between the ID and the keys, as well as the certificates. These certificates have an expiration day to revoke nodes and reduce the possibility of attacks.

As commented in the previous section, every node is responsible for the registration of its pair of keys (KU_n, KS_n) to obtain the certificate, so, it must connect with the CA. Due to the environment where we are, it is possible that some devices have direct connection to the CA, so they have not problems to obtain the certificate to use the application. However, just as the scenario is defined, it is possible that some devices cannot connect with the authority. In this case, the devices that have certificates will act as a certification authority. In order to do this, a node that has not been authenticated by the CA must follow the next steps to guarantee that it is a valid node.

Having a device m without a valid certificate and that wants to install and use the application in a place where it has not access to a certification authority, it will use any other device that belongs to the network as CA to establish communications with the different devices that are legitimate members of the network.

Being n a node with the certificate $Cert_n = (KR_{CA}(KU_n,), KU_{CA})$ and belonging to the network, it is the responsible of acting as CA of the new node m. This node m cannot communicate with anybody because to establish a communication is required a valid certificate. As we explained in the previous section, the node m generates its pair of keys (KU_m, KR_m) and sends it to the node n for its certification. The node n certifies with its private key that has been certificated before by the central CA, being the certificate of the node m the next:

$$Cert_m = (KR_n(KU_m), (KR_{CA}(KU_n), KU_{CA}))$$

With this certificate, a message M signed with the private key of the node m, $KR_m(M)$ can be decrypted by any node of the network thanks to the certificate generated by n, which evidences that it is a valid node. As we can see, in this case a node x that receives this certificate can decrypt a message generated by m and knows that it is valid using the certificate provided by m. When it receives the certificate, it can obtain the public key of m using the public key of the node n, which was authenticated by the CA. Then we know that the node m has been authenticated by n. Therefore, this signature allows us to use the network through a challenge-response scheme with digital signature used in the access control.

The next level could be a node x that only has contact with the node m that was certificated by the node n. The certificate generated by the node m for this node x corresponds to the next chain of certificates:

$$Cert_x = (KR_m(KU_x), (KR_n(KU_m), KU_n), \\ (KR_{AC}(KU_n), KU_{AC})$$

A message M signed by the node x, $KR_x(M)$ can be decrypted by every node of the network following the next steps.

1. KU_n is verified through KU_{AC} applied over $KR_{AC}(KU_n)$
2. KU_m is verified through KU_n applied over $KR_n, (KU_m)$
3. KU_x is verified through KU_m applied over $KR_m(KU_x)$
4. M is verified through KU_x applied over $KR_x(M)$

Therefore, thanks to the concepts of certificate graph and chain of certificates, the use of our application, has been extended (see Figure 4) when the nodes are not registered and authenticated directly by the server and where network infrastructures are not available.

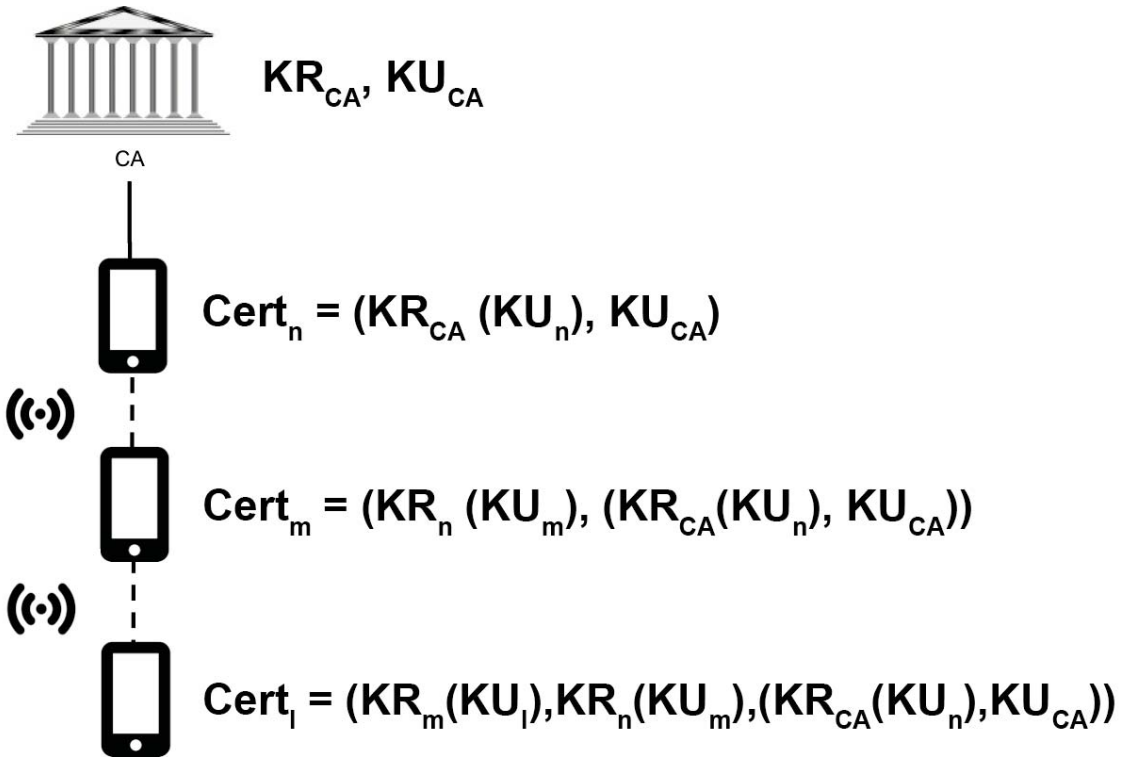


Figure 4: Key Certificates Generation Scheme

5 Conclusions and future work

This work proposes a new communication system for emergency situations. The main objective is to help people that cannot use a communication infrastructure by providing alternative communication channels based on different wireless technologies. The system is based on the use of public chat rooms, with open access for all registered and authenticated people, and private chat rooms that can be used by specific user groups like emergency personnel. Apart of the authentication process based on asymmetric cryptography, the chat rooms are protected with secret key encryption to allow the users to cooperate in the message forwarding in their environment to other users. Moreover, for other users can be registered in cases of not expected emergencies, the certification through chain of certificates in certificate graphs is available.

In the definition of the automatic technology decision system, a study of the battery consumption of the chosen technologies has been done through a benchmark application over the Android system to establish a consumption weight for each technology based on the ratio between different consumptions. During the study of the technologies, several features were identified for each technology, which were taken into account to propose the decision system to choose which technology should be used in every moment, depending on the number of previously discovered peers and other factors.

This work is part of a work in progress so several problems remain open. The security level of the developed application could be increased through its verification by applying different attacks and by testing it with real users. Moreover, we want to establish a battery saving protocol to guarantee the ability to send automatic emergency messages or automatic synchronizing messages with the database, fact that is especially important in emergency situations. Finally, new features could be added, such as voice communication or the possibility of sending files and/or doing videoconferences.

Acknowledgements

This work has been partially supported by the grants TESIS2015010102, TESIS2015010106, IPT-2012-0585- 370000, RTC-2014-1648-8 y TEC2014-54110-R.

References

[1] Alpify [Online], “http://www.alpify.com/es/sistema-de-rescate/”

[2] I. Santos-González, A. Rivero-García, P. Caballero-Gil, C. Hernández-Goya, “Alternative Communication System for Emergency Situations,” International Conference on Web Information Systems and Technologies, pp. 397–402, 2014.

[3] H.C. Jang, Y.N. Lien, T.C. Tsai, "Rescue information system for earthquake disasters based on MANET emergency communication platform," ACM International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly, pp. 623-627, 2009.

[4] W. Stallings, "Cryptography and Network Security:Principles and Practice," Sixth Edition, 2013.

[5] Qualcomm Technologies. LTE Direct Always-on Device-toDevice Proximal Discovery [Online], "https://www.qualcomm.com/media/documents/files/lte-direct-always-on-device-to-device-proximal-discovery.pdf"

[6] S. Capkun, L. Buttyan, J.P. Hubaux, "Self-organized public-key management for mobile ad hoc networks," *IEEE Transactions on Mobile Computing*, no 2.1 pp. 52–64, 2003.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

One-fermion picture for Moshinsky-type atoms and significance of generalized Pauli constraints

Christian Schilling[1]

[1] *Clarendon Laboratory, University of Oxford*

emails: christian.schilling@physics.ox.ac.uk

Abstract

The fermionic exchange symmetry implies restrictions on natural occupation numbers greater than Pauli's famous exclusion principle. First analytic evidence for the significance of those generalized Pauli constraints was found in 2013 for ground states in the form of the quasipinning effect: For Moshinsky-type atoms, i.e. N harmonically coupled fermions confined by a harmonic trap, occupation numbers were found surprisingly close to (but not exactly on) the boundary of the allowed region. We review those findings and discuss the physical relevance of the quasipinning phenomenon. The comprehensive study of the underlying model for different particles, spatial dimensions and coupling strength provided first insights into the mechanism behind quaispinning

Key words: Harmonic interaction, generalized Pauli constraints, natural occupation numbers

1 Introduction and overview of the field

Since its formulation in 1925, Pauli's exclusion principle [1] has played a crucial role in the understanding of various phenomena, such as the atomic structure and related spectral observations, the stability of matter (see e.g. Refs. [2, 3]) and neutron stars. Only one year after its discovery, Heisenberg and Dirac recognized Pauli's exclusion principle to be a consequence of the more substantial fermionic exchange symmetry arising due to the indistinguishability of identical particles [4, 5]. In terms of natural occupation numbers (NONs) λ_i, the eigenvalues of the 1-particle reduced density operator, Pauli's exclusion principle can be stated as

$$0 \leq \lambda_i \leq 1\,, \quad \forall i\,. \tag{1}$$

 ISBN: 978-84-608-6082-2

Here, the NONs are normalized to the particle number N, $\lambda_1 + \ldots + \lambda_d = N$ and we assume that the 1-particle Hilbert space $\mathcal{H}^{(d)}$ is finite, d-dimensional. From a geometrical viewpoint, by ordering the λ_i decreasingly and introducing the λ-vector $\vec{\lambda} \equiv (\lambda_i)_{i=1}^d$ and $\|\vec{x}\|_1 \equiv \sum_{i=1}^d |x_i|$, Eq. (1) restricts such vectors of NONs to the Pauli simplex Σ,

$$\Sigma \equiv \{\vec{\lambda} \in \mathbb{R}^d \mid \|\vec{\lambda}\|_1 = N\,,\, 1 \geq \lambda_1 \geq \ldots \geq \lambda_d \geq 0\}\,. \tag{2}$$

In a number of works [6, 7, 8, 9, 10] the antisymmetry of the N-fermion wave function was found and proven only recently to impose a family of greater restrictions on $\vec{\lambda}$:

$$D_j(\vec{\lambda}) \equiv \kappa_j^{(0)} + \vec{\kappa}_j \cdot \vec{\lambda} \geq 0\,, \quad j = 1, 2, \ldots, r_{N,d}, \tag{3}$$

with $r_{N,d} < \infty$. Note that $(\kappa_j^{(0)}, \vec{\kappa}_j) \in \mathbb{Z}^{d+1}$ as well as the number of constraints $r_{N,d}$ depend on the number of fermions N and the dimension d of the underlying 1-particle Hilbert space. It should be stressed that this recent breakthrough by Klyachko and Altunbulak [8, 7, 9] was part of a more general effort in mathematical physics and quantum information theory [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] addressing the quantum marginal problem. This problem explores and describes the relations between reduced density operators (marginals) of subsystems arising from a common multipartite quantum state. One of the most prominent examples is the (2-body) N-representability problem which is about describing the set of 2-particle reduced density operators being compatible to N-fermion quantum states [24].

The so-called *generalized Pauli constraints* (GPCs) (3) determine a polytope-shaped subset $\mathcal{P}$ (see also Fig. 1),

$$\mathcal{P} \subsetneq \Sigma \subset [0,1]^d\,. \tag{4}$$

In other words, a λ-vector of NONs is compatible to a pure N-fermion quantum state $|\Psi\rangle \in \wedge^N[\mathcal{H}^{(d)}]$ if and only if $\vec{\lambda}$ lies in the polytope $\mathcal{P}$. Here and in the following we typically suppress the dependence of $\mathcal{P}$ and Σ on N, d. To give the reader an idea on how non-trivial the GPC are, we present them for the setting $(N, d) = (3, 6)$ [6]:

$$\lambda_1 + \lambda_6 = \lambda_2 + \lambda_5 = \lambda_3 + \lambda_4 = 1\,, \tag{5}$$

$$D(\vec{\lambda}) \equiv 2 - (\lambda_1 + \lambda_2 + \lambda_4) \geq 0\,. \tag{6}$$

Notice that the inequality $D(\vec{\lambda}) \geq 0$ is manifestly stronger than Pauli's exclusion principle, which just states $2 - (\lambda_1 + \lambda_2) \geq 0$. That some constraints take the form of equalities (instead of inequalities) is specific and happens only for this small setting of three fermions and a 6-dimensional 1-particle Hilbert space.

Given the remarkable result on the GPCs, there is little doubt that these constraints will have some physical relevance as well. For instance, from a general viewpoint, the GPCs may lead to new insights in reduced density matrix functional theory (RDMFT): Usually the minimization of a functional of the 1-particle reduced density operator to determine the ground state is erroneously

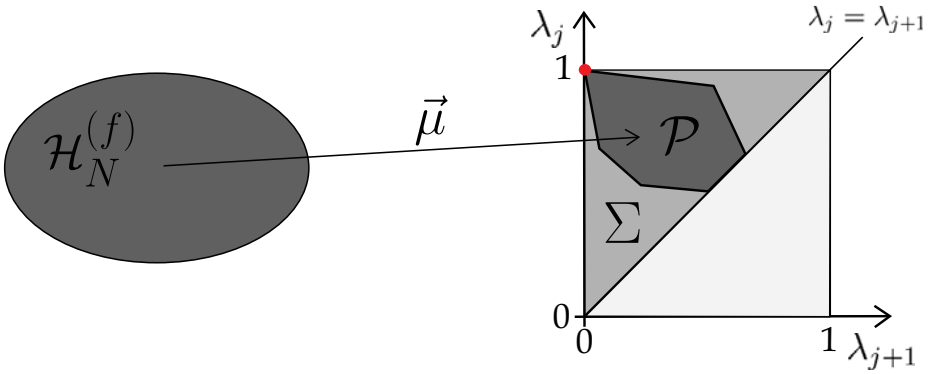


Figure 1: Schematic illustration of the polytope $\mathcal{P}$ of vectors $\vec{\lambda} = (\lambda_i)$ of decreasingly ordered natural occupation numbers defined by the family of generalized Pauli constraints. Only those $\vec{\lambda}$ lying in $\mathcal{P}$ can arise from pure N-fermion quantum states $|\Psi\rangle \in \mathcal{H}_N^{(f)} \equiv \wedge^N[\mathcal{H}_1^{(d)}]$. Hartree-Fock point $\vec{\lambda}_{HF} \equiv (1, \ldots, 1, 0, \ldots)$ is shown as red dot.

considered to be only constrained by (1). Recently, it has been demonstrated for the first time that the GPCs can have a strong influence on the results of the minimization process for several functionals [25]. In addition, the concept of master equations describing the dynamics of NONs may be modified by taking the GPCs into account. The GPCs might be also useful in tomography used for the reconstruction of the 1-particle reduced density matrix given some 1-particle information.

A more specific but potentially quite spectacular relevance of GPCs was postulated by Klyachko [26, 27] in the form of the *pinning* effect: For some systems — from the viewpoint of the 1-particle picture — the ground state minimization process of the energy expectation value $\langle\Psi_N|\hat{H}|\Psi_N\rangle$ for a Hamiltonian $\hat{H}$ might get stuck on the boundary of the polytope $\mathcal{P}$ since any further minimization would violate some GPC (3). Yet, in a first analytic investigation strong evidence was found for *quasipinning* [28]. There, for the ground state of a few-fermion system the NONs were approximately saturating some GPC, $D_j(\vec{\lambda}) \approx 0$, and therefore $\vec{\lambda}$ was found very close to, but not exactly on, the boundary of $\mathcal{P}$.

The occurrence of (quasi-)pinning gives rise to a number of important structural implications, such as a reduced complexity of the N-fermion wave function under expansion in terms of Slater determinants [26, 20, 29] and a constrained dynamical evolution of the system [29, 30].

Over the past few years the study of GPCs, and the search for systems exhibiting (quasi)pinning in particular has therefore become a subject of growing interest among many branches of physics and quantum chemistry [26, 31, 32, 20, 33, 34, 25, 22, 30, 35]. Yet most of those works resorted to numerical methods, and in addition employed quite strong approximations: The 1-particle Hilbert space was truncated from infinite dimensions to at most six up to eight. As an unfortunate consequence, the NONs of the approximated ground states turn out to differ quite a lot from those of the *correct* ground state and no conclusive statement on the occurrence of (quasi)pinning for the correct

ground state was possible [36]. This also renews the caveat already expressed in [28]: *"it is likely extremely challenging to use numerical methods to distinguish between genuinely pinned and mere quasipinned states. This underscores the need for analytical analyses,..."*. Moreover, little has been understood so far about the origin of (quasi-)pinning.

In the following, we review Ref. [28] which has opened this new field and we will present the follow up results [35, 37] as well.

2 The model and an effective 1-particle picture

The system we are considering consists of N identical (here fully polarized/spinless) particles of mass m that are confined in an n-dimensional (not necessarily isotropic) harmonic trapping potential characterized by its trapping frequencies $\{\omega^{(\alpha)}\}_{\alpha=1}^{n}$. In addition to the external potential, a harmonic particle-particle interaction of strength K will be taken into account. Consequently, the Hamiltonian reads

$$H_N = \sum_{i=1}^{N}\left(\frac{\vec{p}_i^{\,2}}{2m} + \frac{m}{2}\vec{x}_i^{\,t}\Omega\vec{x}_i\right) + \frac{K}{2}\sum_{1\leq i<j\leq N}(\vec{x}_i-\vec{x}_j)^2, \tag{7}$$

where $\vec{p}_i = (p_i^{(\alpha)})_{\alpha=1}^n$ and $\vec{x}_i = (x_i^{(\alpha)})_{\alpha=1}^n$ represent the momentum and position operators of the i-th particle and $\Omega \equiv \mathrm{diag}(\omega_1^2,\ldots,\omega_n^2)$.

A priori, the Hamiltonian (7) acts as an operator on the N-particle Hilbert space $\mathcal{H}^{(N)} = \bigotimes_{i=1}^{N}\mathcal{H}$, where the 1-particle Hilbert space $\mathcal{H}$ is given by $\mathcal{H} = L^2(\mathbb{R}^n)$. Any permutation of particles leaves H_N invariant. In particular, this allows us to treat the N particles as indistinguishable fermions and thus restrict the Hamiltonian (7) to the subspace

$$\mathcal{H}_N^{(f)} \equiv \wedge^N[\mathcal{H}] = \mathcal{A}_N\mathcal{H}^{(N)} \lneq \mathcal{H}^{(N)} \equiv \mathcal{H}^{\otimes^N} \tag{8}$$

of antisymmetric states. Here, $\mathcal{A}_N$ represents the antisymmetrising operator. In order to derive the set of fermionic eigenstates of the Hamiltonian (7) we therefore initially may derive the set of all N-particle eigenstates followed by a projection onto $\mathcal{A}_N\mathcal{H}^{(N)}$.

For interacting fermions, in contrast to non-interacting fermions, one cannot expect that the structure of the energy eigenstates can be elegantly described by exploiting the elementary and convenient 1-fermion picture. Yet, a bit surprisingly, this is still possible at least for the ground state of Hamiltonian (7):

Theorem 2.1. *The N-fermion ground state Ψ of the model (7) is given by*

$$\Psi(\vec{x}_1,\ldots,\vec{x}_N) = \mathcal{N}\begin{vmatrix} \phi_{\boldsymbol{\mu}_1}^{(\tilde{\boldsymbol{l}})}(\vec{x}_1) & \cdots & \phi_{\boldsymbol{\mu}_1}^{(\tilde{\boldsymbol{l}})}(\vec{x}_N) \\ \vdots & & \vdots \\ \phi_{\boldsymbol{\mu}_N}^{(\tilde{\boldsymbol{l}})}(\vec{x}_1) & \cdots & \phi_{\boldsymbol{\mu}_N}^{(\tilde{\boldsymbol{l}})}(\vec{x}_N) \end{vmatrix} \times e^{\vec{X}^t\boldsymbol{B}\vec{X}}, \tag{9}$$

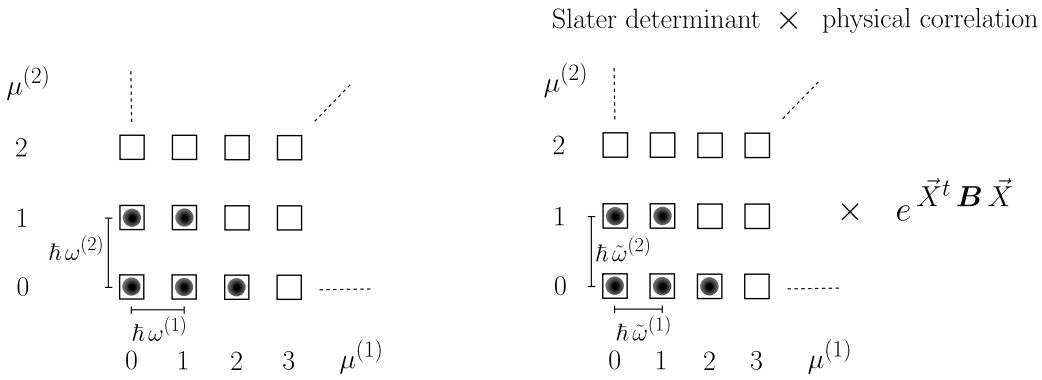


Figure 2: Graphical illustration of the fermionic ground state $|\Psi\rangle$ for the exemplary case of $n = 2$ and $N = 5$. For zero interaction (left) $|\Psi\rangle$ is given by a single Slater determinant obtained by filling the N 'boxes' with lowest energy respecting Pauli's exclusion principle. Each box, labeled by $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ describes a 1-particle orbital given by a harmonic oscillator state in n dimensions with corresponding frequencies $\omega^{(\alpha)}$, $\alpha = 1, \ldots, n$. For finite interaction (right) this Slater determinant structure is modified by only a 'correlation term' $e^{\vec{X}^t B \vec{X}}$ for the center of mass $\vec{X} \equiv \frac{1}{N}(\vec{x_1} + \ldots \vec{x}_N)$ (and the relative distances between the boxes change).

where we introduced the center of mass vector $\vec{X} \equiv \frac{1}{N}(\vec{x}_1 + \ldots \vec{x}_N)$, $\boldsymbol{B} \equiv diag(B^{(1)}, \ldots, B^{(n)})$, $B^{(\alpha)} \equiv N\frac{(l^{(\alpha)})^2 - (\tilde{l}^{(\alpha)})^2}{2(l^{(\alpha)}\tilde{l}^{(\alpha)})^2}$, $\tilde{l}^{(\alpha)} \equiv \sqrt{\frac{\hbar}{m\omega^{(\alpha)}\sqrt{1+NK/(m(\omega^{(\alpha)})^2)}}}$, $\phi^{(\boldsymbol{l})}_{\boldsymbol{\mu}}$ *the* $\boldsymbol{\mu}$*-th Hermite function in* n *dimensions with length scale vector* $\boldsymbol{l}$ *and* $\mathcal{N}$ *is a normalisation constant. The quantum number vectors* $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ *in* (9) *are chosen such that the following energy function*

$$E_{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_N} \equiv \sum_{i=1}^{N} \tilde{\varepsilon}_{\boldsymbol{\mu}_i} \tag{10}$$

is minimal, yet respecting Pauli's exclusion principle[1] *(i.e. all* $\boldsymbol{\mu}_i$ *are different) and* $\tilde{\varepsilon}_{\boldsymbol{\mu}_i} \equiv \sum_{\alpha=1}^{n}(\mu^{(\alpha)} + \frac{1}{2})\,\hbar\tilde{\omega}^{(\alpha)}$.

The proof of Theorem 2.1 is less trivial and can be found in Ref. [37]. Theorem 2.1 and the structure of the ground state are illustrated in Fig. 2.

As a caveat, we would like to also stress that the excited states are *not* given by filling 'boxes' of higher energy and then multiplying the corresponding Slater determinant by the exponential factor as in Eq. (9). Indeed, their structure is more complicated and the single Slater determinant in Eq. (9)

[1]Strictly speaking, this is not exactly the Pauli exclusion principle since it is not applied to the total quantum state Ψ but just to one of its two factors. In a similar way, one should understand the term 'distributing N fermions in N shells' more as an analogy.

 ISBN: 978-84-608-6082-2

would need to be replaced by a linear combination of several Slater determinants, expressing the additional correlations in the system.

3 (Quasi)pinning analysis

The ground state of the Hamiltonian (7) can easily be determined for every particle number and spatial dimension according to Theorem 2.1. To explore the significance of the GPC we need to determine its natural occupation numbers (NONs). In a first step, as an exercise in Gauusian integration we can determine the 1-particle reduced density operator. By using either a perturbational approach for weak couplings or an exact numerical approach for a fixed coupling we can determine the corresponding NONs. Given such vectors $\vec{\lambda}$ at hand we can explore the occurrence of (quasi)pinning. The analysis of the distance to the polytope boundary is quite subtle. The underlying 1-particle Hilbert space is infinite-dimensional, but the polytopes are known only for dimensions $d \leq 10$. However, as justified by the 'concept of truncation' [20, 35] various NONs sufficiently close to 0 can be neglected and the occurrence of (quasi)pinning can be explored in the 'truncated setting' of the remaining NONs. The smallest distance of the truncated vector to the polytope boundary translates to the same distance of the full vector to the boundary of the total polytope up to a small truncation error given by the largest neglected NON.

For instance, for the case of just three fermions in one spatial dimension as studied in Ref. [28] the vector of NONs was truncated to just the first seven NONs,

$$\begin{aligned}
1-\lambda_1 &= \frac{40}{729}\delta^6 - \frac{1390}{59049}\delta^8 + o(\delta^{10})\,, & \lambda_5 &= \frac{2}{9}\delta^4 - \frac{232}{729}\delta^6 + \frac{3976}{10935}\delta^8 + o(\delta^{10}) \\
1-\lambda_2 &= \frac{2}{9}\delta^4 - \frac{232}{729}\delta^6 + \frac{3926}{10935}\delta^8 + o(\delta^{10})\,, & \lambda_6 &= \frac{40}{729}\delta^6 - \frac{2200}{59049}\delta^8 + o(\delta^{10}) \\
1-\lambda_3 &= \frac{2}{9}\delta^4 - \frac{64}{243}\delta^6 + \frac{81902}{295245}\delta^8 + o(\delta^{10})\,, & \lambda_7 &= \frac{80}{2187}\delta^8 + o(\delta^{10}) \\
\lambda_4 &= \frac{2}{9}\delta^4 - \frac{64}{243}\delta^6 + \frac{73802}{295245}\delta^8 + o(\delta^{10})\,, & \lambda_8 &= o(\delta^{10})
\end{aligned} \tag{11}$$

where we introduced the dimensionless coupling parameter $\delta \equiv \log\left(1 + \frac{3K}{4m\omega^2}\right)$. The (quasi)pinning analysis of this truncated vector (11) leads to a minimal distance to the polytope boundary [28]

$$D(\delta) \sim const \times \delta^8\,. \tag{12}$$

This quasipining is non-trivial in the sense that the distance (12) to the polytope boundary ∂P is by four orders in the coupling δ smaller than the distance to the Hartree-Fock point (lies on the boundary, see Figure 1), which behaves as δ^4. Moreover, quasipinning is not only present in the regime of week interaction ($|\delta|$ small), but also for medium interaction strengths. For instance, for the harmonic analogue of the Lithium atom, $\frac{3K}{m\omega^2} = \frac{1}{3}$, we found $D = 5.8 \cdot 10^{-8}$. Remarkably, as it was found in Ref. [35] the quasipinning even increases by loading the 1-dimensional trap with more

particles,

$$D_N(\delta) \sim const \times \delta^{2N} \,, \tag{13}$$

for $N \geq 4$. This increase of the quasipinning strength suggests the existence of a 'Pauli pressure', pressing the vector $\vec{\lambda}$ of NONs further to the boundary of the polytope when more fermions are added to the trap.

In general, it turns out that the quasipinning becomes weaker in higher spatial dimensions. For details we refer the reader to Ref. [37]. This again provides evidence for a 'Pauli pressure' since the conflict of energy minimization and antisymmetry in higher spatial dimensions is reduced due to degeneracies associated with the additional angular degrees of freedom.

Acknowledgements

We thank F. Tennie and V. Vedral for helpful discussions. We gratefully acknowledge financial support from the Swiss National Science Foundation (Grant P2EZP2 152190) and the Oxford Martin Programme on Bio-Inspired Quantum Technologies (CS).

References

[1] W. Pauli. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Z. Phys.*, 31:765–783, 1925.

[2] F. J. Dyson and A. Lenard. Stability of matter. *IS J. Math. Phys.*, (8):423–434, 1967.

[3] Elliott H. Lieb. The stability of matter: from atoms to stars. *Bull. Amer. Math. Soc. (N.S.)*, 22(1):1–49, 01 1990.

[4] P. A. M. Dirac. On the theory of quantum mechanics. *Proc. R. Soc. A*, 112(762):661–677, 1926.

[5] W. Heisenberg. Mehrkörperproblem und Resonanz in der Quantenmechanik. *Z. Phys.*, 38:411–426, 1926.

[6] R.E. Borland and K. Dennis. The conditions on the one-matrix for three-body fermion wavefunctions with one-rank equal to six. *J. Phys. B*, 5(1):7, 1972.

[7] A. Klyachko. Quantum marginal problem and N-representability. *J. Phys.*, 36(1):72, 2006.

[8] M. Altunbulak and A. Klyachko. The Pauli principle revisited. *Commun. Math. Phys.*, 282:287–322, 2008.

[9] M. Altunbulak. *The Pauli principle, representation theory, and geometry of flag varieties*. PhD thesis, Bilkent University, 2008.

[10] M. B. Ruskai. Connecting N-representability to Weyl's problem: the one-particle density matrix for N = 3 and R = 6. *J. Phys. A*, 40(45):F961, 2007.

[11] C. Müller. Sufficient conditions for pure state N-representability. *J. Phys. A*, 32(22):4139, 1999.

[12] A. Higuchi, A. Sudbery, and J. Szulc. One-qubit reduced states of a pure many-qubit state: polygon inequalities. *Phys. Rev. Lett.*, 90, 2003.

[13] S. Bravyi. Requirements for compatibility between local and multipartite quantum states. *Quant. Inf. Comp.*, 4:012, 2004.

[14] A. Klyachko. Quantum marginal problem and representations of the symmetric group. *arXiv:quant-ph/0409113*, September 2004.

[15] S. Daftuar and P. Hayden. Quantum state transformations and the Schubert calculus. *Ann. Phys.*, 315(1):80 – 122, 2005.

[16] M. Christandl and G. Mitchison. The spectra of quantum states and the Kronecker coefficients of the symmetric group. *Commun. Math. Phys.*, 261:789–797, 2006.

[17] Y.-K. Liu, M. Christandl, and F. Verstraete. Quantum computational complexity of the N-representability problem: QMA complete. *Phys. Rev. Lett.*, 98:110503, 2007.

[18] W. Hall. Compatibility of subsystem states and convex geometry. *Phys. Rev. A*, 75:032102, 2007.

[19] J. Eisert, T. Tyc, T. Rudolph, and B.C. Sanders. Gaussian quantum marginal problem. *Comm. Math. Phys*, 280(1):263–280, 2008.

[20] C. Schilling. *Quantum marginal problem and its physical relevance*. PhD thesis, ETH-Zürich, 2014.

[21] M. Krbek, T. Tyc, and J. Vlach. Inequalities for quantum marginal problems with continuous variables. *J. Math. Phys.*, 55(6), 2014.

[22] A. Lopes. *Pure univariate quantum marginals and electronic transport properties of geometrically frustrated systems*. PhD thesis, University of Freiburg, 2015.

[23] T. Tyc and J. Vlach. Quantum marginal problems. *Eur. Phys. J. D*, 69(9), 2015.

[24] A. J. Coleman and V. I. Yukalov. *Reduced Density Matrices: Coulsons Challenge*. Springer, New York, 2000.

[25] I. Theophilou, N.N. Lathiotakis, M. Marques, and N. Helbig. Generalized Pauli constraints in reduced density matrix functional theory. *J. Chem. Phys.*, 142(15), 2015.

[26] A. Klyachko. The Pauli exclusion principle and beyond. *arXiv:0904.2009*, 2009.

[27] A. Klyachko. The Pauli principle and magnetism. *arXiv:1311.5999*, 2013.

[28] C. Schilling, D. Gross, and M. Christandl. Pinning of fermionic occupation numbers. *Phys. Rev. Lett.*, 110:040404, Jan 2013.

[29] C. Schilling. *The quantum marginal problem*, chapter -1, pages 165–176.

[30] C. Schilling. Hubbard model: Pinning of occupation numbers and role of symmetries. *Phys. Rev. B*, 92:155149, Oct 2015.

[31] C. Benavides-Riveros, J. Gracia-Bondia, and M. Springborg. Quasipinning and entanglement in the lithium isoelectronic series. *Phys. Rev. A*, 88:022508, 2013.

[32] R. Chakraborty and D.A. Mazziotti. Generalized Pauli conditions on the spectra of one-electron reduced density matrices of atoms and molecules. *Phys. Rev. A*, 89:042505, 2014.

[33] C. Schilling. Quasipinning and its relevance for N-fermion quantum states. *Phys. Rev. A*, 91:022105, Feb 2015.

[34] C. L. Benavides-Riveros and M. Springborg. Quasipinning and selection rules for excitations in atoms and molecules. *Phys. Rev. A*, 92:012512, Jul 2015.

[35] F. Tennie, D. Ebler, V. Vedral, and C. Schilling. Pinning of fermionic occupation numbers: General concepts and one spatial dimension. *Phys. Rev. A*, 93:042126, 2016.

[36] S. Knecht and M. Reiher. private communication.

[37] F. Tennie, V. Vedral, and C. Schilling. Pinning of Fermionic Occupation Numbers: Higher Spatial Dimensions and Spin. *in preparation*, 2016.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Ab initio calculation of electronically excited states for large molecular systems

Martin Schütz[1]

[1] *Institute for Physical and Theoretical Chemistry, University of Regensburg, Germany*

emails: martin.schuetz@chemie.uni-regensburg.de

Abstract

When a molecular system absorbs a photon of a certain energy it is transferred to the corresponding (energetically matching) electronically excited state. Within the Born-Oppenheimer approximation the system is "beamed" into a new energy landscape, which may look entirely different than that of the electronic ground state. New reaction channels open, and the system may have multiple ways to shed the energy acquired by photon absorption.

Electronically excited states therefore play a major role in many physical, chemical, and biological processes – whenever photons are absorbed they are of relevance. For that reason it is desirable to have a toolbox of first-principle (or *ab initio*) methods at hand, with which it is possible to compute reliably properties of excited states, such as excitation energies, changes in the electronic density, transition strength, and nuclear gradients. In the past decade efficient implementations of coupled cluster linear response (CCLR) became available, which are able to treat systems of a size previously only accessible to time-dependent density functional theory (TD-DFT). The latter has many shortcomings and is qualitatively wrong for systems with pronounced charge transfer character.

In the Regensburg theory group we have developed over the years efficient CCLR methods based on *localized* molecular orbitals. This allows it to exploit the locality of dynamic correlation effects und to reduce the computational complexity, *i.e.*, the scaling of the computational cost with molecular size. These local CCLR methods enable calculations of properties of singlet and triplet excited states including excitation energies [1, 2], orbital un-relaxed and relaxed first-order properties [3, 4, 5, 6], and analytic gradients w.r. to nuclear displacements [7, 8]. Quite recently, we extended the methodology from (number conserving) excitations to (not number conserving) ionization potentials [9]. The latter may also offer a convenient pathway to excited states of large open-shell radicals (via differences of ionization potentials). Radicals are important intermediates in many chemical reactions and can be detected by means of spectroscopical methods. In my contribution I will report on these recent developments.

  ISBN: 978-84-608-6082-2

References

[1] D. Kats, T. Korona, and M. Schütz, *J. Chem. Phys.* **125**, 104106 (2006).

[2] D. Kats and M. Schütz, *J. Chem. Phys.* **131**, 124117 (2009).

[3] D. Kats, T. Korona, and M. Schütz, *J. Chem. Phys.* **127**, 064107 (2007).

[4] D. Kats and M. Schütz, *Z. Phys. Chem.* **224**, 601 (2010).

[5] K. Freundorfer, D. Kats, T. Korona, and M. Schütz, *J. Chem. Phys.* **133**, 244110 (2010).

[6] K. Ledermüller, D. Kats, and M. Schütz, *J. Chem. Phys.* **139**, 084111 (2013).

[7] K. Ledermüller and M. Schütz, *J. Chem. Phys.* **140**, 164113 (2014).

[8] M. Schütz, *J. Chem. Phys.* **142**, 214103 (2015).

[9] G. Wälz, D. Usvyat, T. Korona, and M. Schütz, *J. Chem. Phys.* **144**, 084117 (2016).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

The Extended Lennard-Jones Potential for Cubic Solids

Peter Schwerdtfeger[1] and Elke Pahl[2]

[1] *Centre for Theoretical Chemistry and Physics, The New Zealand Institute for Advanced Study (NZIAS), Massey University Albany, Private Bag 102904, Auckland 0745, New Zealand*

[2] *Centre for Theoretical Chemistry and Physics, Institute for Natural and Mathematical Sciences (INMS), Massey University Albany, Private Bag 102904, Auckland 0745, New Zealand*

emails: p.a.schwerdtfeger@massey.ac.nz, e.pahl@massey.ac.nz

Abstract

The Lennard-Jones (LJ) potential is the most widely used interaction potential between atoms with widespread applications in physical, chemical and biological sciences. This simple potential also has the advantage that the cohesive energy, pressure and the bulk modulus of a simple solid (simple cubic, body-centered cubic, and face-centered cubic) can be expressed analytically as a function of volume using the Lennard-Jones-Ingham coefficients derived from infinite series expansions. In a similar procedure to Lennard-Jones we derive analytical expressions for the zero-point vibrational energy and first-order anharmonicity corrections for these crystals by an inverse power expansion in terms of the internuclear distance, which we call the Extended Lennard-Jones potential. These new expressions are applied to the Lennard-Jones potentials for the rare gas solids from helium to krypton.

Key words: Lennard-Jones Potential, Lennard-Jones-Ingham Coefficients, Solid State Properties, Rare Gases

  ISBN: 978-84-608-6082-2

Introduction

The Lennard-Jones potential (LJ) is a simple but computer efficient interaction potential between atoms and has the following form,[1, 2]

$$V_{\mathrm{LJ}}(r) = \epsilon \left[(r_e/r)^{12} - 2\,(r_e/r)^{6} \right] \tag{1}$$

This potential form has widespread applications in the simulation of atoms and molecules in the gas, liquid and the solid phase.[3] It contains only two parameters, the equilibrium distance r_e and the binding energy ϵ. We have recently extended the simple LJ potential into a more general inverse power series,[4]

$$V_{\mathrm{ELJ}}(r) = \sum_{n>3} c_n r^{-n} \tag{2}$$

with the correct boundary conditions for the coefficients c_n such that $V_{\mathrm{ELJ}}(r_e) = -\epsilon$ at the equilibrium distance r_e of the potential energy curve. This so-called extended LJ potential (ELJ) is more accurate compared to the simple LJ potential, and has the advantage of being still computationally efficient and reasonably accurate when compared with other analytical forms treating the short- and long-range behaviour of the interaction potential separately.[5, 6, 7] It was used recently in the simulation of solid-to-liquid phase transitions for neon and argon.[8, 9, 10, 11]

Another advantage of the ELJ form (2) is that one can find an analytical expression for the cohesive energy (per atom) $E_{\mathrm{ELJ}}^{\mathrm{coh}}$ for the sc (simple cubic), bcc, fcc and hcp crystals in terms of the Lennard-Jones-Ingham (LJI) coefficients L_n,[4, 12, 13]

$$E_{\mathrm{ELJ}}^{\mathrm{coh}}(r_s) = \frac{1}{2} \sum_{n>3} c_n L_n r_s^{-n} \tag{3}$$

Here r_s is the nearest neighbour distance in the solid. The LJI coefficients for these structures have been obtained recently to high precision,[4]

$$L_m^{\mathrm{sc}} = \sum_{i,j,k \in \mathbb{Z} \setminus (0,0,0)} \left(i^2 + j^2 + k^2 \right)^{-\frac{m}{2}} \tag{4}$$

$$L_m^{\mathrm{bcc}} = \left(\frac{\sqrt{3}}{2} \right)^m \left(L_m^{\mathrm{sc}} + \sum_{i,j,k \in \mathbb{Z}} \left(\left(i + \tfrac{1}{2}\right)^2 + \left(j + \tfrac{1}{2}\right)^2 + \left(k + \tfrac{1}{2}\right)^2 \right)^{-\frac{m}{2}} \right) \tag{5}$$

$$L_m^{\mathrm{fcc}} = 3 \sum_{i,j,k \in \mathbb{Z} \setminus (0,0,0)} \left(2i^2 + j^2 + k^2 \right)^{-\frac{m}{2}} - 2^{1-\frac{\mathrm{m}}{2}} L_m^{\mathrm{sc}} \tag{6}$$

$$L_m^{\rm hcp} = \sum_{i,j,k\in Z\setminus(0,0,0)} \left(i^2+j^2+\tfrac{8}{3}k^2\right)^{-\frac{m}{2}} + \sum_{i,j,k\in\mathbb{Z}} \left(\left(i+\tfrac{1}{3}\right)^2+\left(j+\tfrac{1}{3}\right)^2+\left(i+\tfrac{1}{3}\right)\left(j+\tfrac{1}{3}\right)+\tfrac{8}{3}\left(k+\tfrac{1}{2}\right)^2\right)^{-\frac{m}{2}} \quad (7)$$

For these solids one can re-express eq.(3) in terms of the volume to obtain the corresponding analytical expressions for the pressure $P = -\partial E/\partial V$ and the bulk modulus $B = V\partial^2 E/\partial V^2$ of a solid. We have now been successful to derive simple expressions for the zero-point vibrational energy and corresponding anharmonicity correction similar to eq.(3), and apply these expressions to the rare gas solids from helium to krypton.

Results

The results using a simple (6-12) LJ potential for our analytical expressions are shown in Table 1. The total energies $E_{\rm total}$ show that both isotopes of solid helium He3 and He-4 are unbound within this approximation. However, it is a well known fact that helium becomes solid only under pressure. For the heavier rare gases experimental cohesive energies are available,[16, 17] which are (in [μHa]) for -1002 for neon, -3268 for argon and -4502 for krypton. The rather large differences are due to the simple LJ form and to the fact that higher than 2-body terms are neglected, which has been considered by Stoll and co-workers.[18] Extensions to the more accurate ELJ form including higher n-body forces are currently underway in our research group.

Acknowledgements

This work has been supported by the Marsden fund administered by the Royal Society of New Zealand.

References

[1] J. E. Jones, *On the determination of molecular fields. II. From the equation of state of a gas*, Proc. R. Soc. Lond. A **106** (1924) 463–477.

[2] J. E. Lennard-Jones, *Cohesion*, Proceedings of the Physical Society **43**, 5 (1931) 461.

[3] J. O. Hirschfelder, C. F. Curtiss, and R. B. Bird, , *The Molecular Theory of Gases and Liquids*, John Wiley & Sons, New York, 1964.

Table 1: Lennard Jones (6-12) parameters for the fcc lattices of the rare gases. Binding energies -ϵ, cohesive energies $E_{\rm coh}$, zero-point vibrational energies (ZPVE) $\Delta E_{\rm ZPVE}$, anharmonicity corrections $\Delta E_{\rm AZPVE}$ and total cohesive energies $E_{\rm total} = E_{\rm coh} + \Delta E_{\rm ZPVE} + \Delta E_{\rm AZPVE}$ in [μHa] at $r_s^{\rm min}$. Equilibrium distances r_e of the diatomic, nearest neighbour distance of the solid $r_s^{\rm min}$, ZPVE corrected nearest neighbour distance $r_s^{\rm ZPVE}$ in [Å]. Atomic masses m are in [u]. Binding energies and equilibrium distances are taken from Refs.[6, 14, 15].

Element	m	-ϵ	$E_{\rm coh}$	$\Delta E_{\rm ZPVE}$	$\Delta E_{\rm AZPVE}$	$\Delta E_{\rm total}$	r_e	$r_s^{\rm min}$	$r_s^{\rm ZPVE}$
^{3}He	3.016	-34.8	-299.8	462.3	153.2	315.6	2.968	2.882	-
^{4}He	4.003	-34.8	-299.8	401.3	115.4	216.9	2.968	2.882	-
^{20}Ne	19.992	-133.5	-1149.4	337.7	21.3	-790.4	3.090	3.001	3.125
^{40}Ar	39.962	-453.2	-3902.5	361.4	7.2	-3533.9	3.762	3.654	3.698
^{84}Kr	83.912	-636.1	-5477.3	276.8	3.0	-5197.5	4.016	3.900	3.926

[4] P. Schwerdtfeger, N. Gaston, R. P. Krawczyk, R. Tonner, and G. E. Moyano, *Extension of the Lennard-Jones potential: Theoretical investigations into rare-gas clusters and crystal lattices of He, Ne, Ar and Kr using many-body interaction expansions*, Phys. Rev. B **73** (2006) 064112–1–19.

[5] K. T. Tang, and J. P. Toennies, *An improved simple model for the van der waals potential based on universal damping functions for the dispersion coefficients*, J. Chem. Phys. **80** (1984) 3726–3741.

[6] B. Jäger, R. Hellmann, E. Bich, E. and Vogel, *Ab initio pair potential energy curve for the argon atom pair and thermophysical properties of the dilute argon gas. I. Argon–argon interatomic potential and rovibrational spectra*, Mol. Phys. **107** (2009) 2181–2188.

[7] K. Patkowski, and K. Szalewicz, *Argon pair potential at basis set and excitation limits*, J. Chem. Phys. **133** (2010) 094304–1–20.

[8] E. Pahl, F. Calvo, L. Koči, and P. Schwerdtfeger, *Accurate Melting Temperatures for Neon and Argon from Ab Initio Monte Carlo Simulations*, Angew. Chem. Int. Ed. **47** (2008) 8207–8210.

[9] E. Pahl, F. Calvo, and P. Schwerdtfeger, *The Importance of Accurate Interaction Potentials in the Melting of Argon Nanoclusters*, Int. J. Quant. Chem. 109 **9** (2009) 1812–1819.

 ISBN: 978-84-608-6082-2

[10] F. Senn, J. Wiebke, O. Schumann, S. Gohr, P. Schwerdtfeger, and E. Pahl, *Melting of non-magic argon clusters and extrapolation to the bulk limit*, J. Chem. Phys. **140** (2014) 044325–1–5.

[11] J. Wiebke, E. Pahl, and P. Schwerdtfeger, *Melting at high pressure: Can first-principles computational chemistry challenge diamond-anvil cell experiments?*, Angew. Chem. Int. Ed. **52** (2013) 13202–13205.

[12] J. E. Jones, *On the determination of molecular fields. III. From crystal measurements and kinetic theory data*, Proc. Royal Soc. London A: Math. Phys. Engin. Sci. **106** (1924) 709–718.

[13] J. E. Jones and A. E. Ingham, *On the calculation of certain crystal potential constants, and on the cubic crystal of least potential energy*, Proc. Royal Soc. London A: Math. Phys. Engin. Sci. **107** (1925) 636–653.

[14] E. Bich, R. Hellmann, and E. Vogel, *Ab initio potential energy curve for the neon atom pair and thermophysical properties for the dilute neon gas. II. Thermophysical properties for low-density neon*, Mol. Phys. **106** (2008) 1107–1122.

[15] W. Cencek, M. Przybytek, J. Komasa, J. B. Mehl, B. Jeziorski, and K. Szalewicz, *Effects of adiabatic, relativistic, and quantum electrodynamics interactions on the pair potential and thermophysical properties of helium*, J. Chem. Phys. **136** (2012) 224303–1–31.

[16] G. T. McConville, *New values of sublimation energy l0 for natural neon and its isotopes*, J. Chem. Phys. **60** (1974) 4093–4093.

[17] L. A. Schwalbe, R. K. Crawford, H. H. Chen, and R. A. Aziz *Thermodynamic consistency of vapor pressure and calorimetric data for argon, krypton, and xenon*, J. Chem. Phys. **66** (1977) 4493–4502.

[18] K. Rościszewski, B. Paulus, P. Fulde, and H. Stoll, *Ab initio calculation of ground-state properties of rare-gas crystals*, Phys. Rev. B **60** (1999) 7905–7910.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

DRBEM Solution of Biomagnetic Fluid Flow under a Point Source Magnetic Field

P. Senel[1] and M. Tezer-Sezgin[1]

[1] *Department of Mathematics, Middle East Technical University, 06800, Ankara, Turkey*

emails: psenel@metu.edu.tr, munt@metu.edu.tr

Abstract

In this paper, we investigate the fully developed, laminar flow of an electrically non-conducting, viscous, biomagnetic fluid in a long impermeable pipe. The flow is in the $z-$direction (axis of the pipe) and generated by a given constant pressure gradient making the problem two-dimensional in the cross-section of the pipe (cavity). It is also under the influence of a point magnetic source which results from a magnetic wire placed below and passing through the pipe. The gravitational force and the temperature difference between the walls of the cavity create the forced convection flow on the biomagnetic fluid. We use the Dual Reciprocity Boundary Element Method (DRBEM) for solving the governing equations by taking all the terms other than Laplacian as inhomogeneity in the Poisson's equations for the velocity components, pressure and the temperature of the fluid. The fundamental solution of Laplace equation is made use of converting differential equations to boundary integral equations. We discretize only the boundary of the problem with constant elements and use sufficient number of interior nodes which reduces the computational cost. The unknown pressure boundary conditions are approximated through momentum equations by using finite difference approximation for the pressure gradients at the boundary and the interior nodes. All the space derivatives are calculated by DRBEM coordinate matrix. Pipe axis velocity is also computed. The effects of magnetization and the buoyancy force on the fluid with or without viscous dissipation term in the energy equation are investigated. As magnetic field intensity Mn increases the isotherms are almost symmetrically divided through the hot and cold walls. The flow in the cavity accelerates and axial velocity around the source decelerates. For high Rayleigh number $Ra = 10^5$ the buoyancy force effect becomes dominant and it changes the flow behavior in the cavity totally even when $Mn = 10$. The viscous dissipation effect is observed only when both magnetic and buoyancy forces are high ($Mn = 80$, $Ra = 10^5$). It retards the flow and reduces the heat transfer in the cavity.
Key words: DRBEM, biomagnetic fluid, forced convection

 ISBN: 978-84-608-6082-2

1 Introduction

The study of biomagnetic fluid flow under the effect of a magnetic field has many medical applications such as reduction of blood flow during surgeries, magnet therapies, drug targeting and the oxygenation of cells. The governing equations for the flow of incompressible biomagnetic fluids are similar to those used in Ferrohydrodynamics (FHD) which deals with electrically non-conducting fluids and the body force is due to the gravitational and polarization force [1]. The most characteristic biomagnetic fluid is blood which is considered as a magnetic fluid because of the hemoglobin molecule, a form of iron oxides, carried by the red blood cells. Solution of such flows are given by Tzirtzilakis et al. [2] in a 3D rectangular duct by using pressure-linked pseudotransient method on a common grid. Loukopoulos and Tzirtzilakis [3] studied the influence of the spatially varying magnetic field on the biomagnetic fluid flow in a channel by developing a numerical technique based on finite differences. They assumed the magnetization of the fluid is varying linearly with the temperature and the magnetic field strength. The effect of gravitational acceleration on unsteady biomagnetic fluid flow in a channel under the influence of a spatially varying magnetic field is investigated by Idris et al. [4]. They have used pressure correction method with SIMPLE algorithm.
In this study, we investigate the effect of a non-uniform magnetic field and the buoyancy force on steady, fully developed, laminar, forced convection flow of viscous, incompressible, non-conducting, magnetizable biomagnetic fluid (blood) on the 2D square cross-section of a long impermeable pipe. The governing equations in terms of velocity, pressure and the temperature of the fluid are solved iteratively by DRBEM and the boundary of the square cavity is discretized by constant elements. Depending on the variations of $u-$ and $v-$velocities, axial velocity component is also solved. Stream function equation is introduced in order to visualize the flow in the cavity. The equation for pressure is obtained by using the momentum and the continuity equations. The numerical results are given in terms of velocity, pressure contours, isotherms and the streamlines of the fluid for increasing Magnetic number and Rayleigh number values neglecting the viscous dissipation term in the energy equation. The effect of the viscous dissipation on the numerical results is also discussed.

2 Governing Equations

The forced convection steady flow of an electrically non-conducting biomagnetic fluid is considered in the 2D transverse plane which is the square cross-section of the pipe. Flow is subjected to an applied magnetic point source placed below the pipe on the symmetry axis. The flow in the axial direction is developed by a given constant pressure along the $z-$axis. Being a fully developed flow the pressure is split into two parts as in [5]

$$P(x, y, z) = p(x, y) + P_1(z) \tag{1}$$

$$\frac{\partial P}{\partial z} = \frac{\partial P_1}{\partial z} = P_z = constant\ . \tag{2}$$

The magnetic field strength H is given by [2]

$$H(x,y) = \frac{|b|}{\sqrt{(x-a)^2 + (y-b)^2}} \tag{3}$$

where (a,b) denotes the place of the point source, (x,y) is any point inside the cavity. Then, the continuity equation, equations of motion and the energy equation in non-dimensional form in terms of velocity (u,v,w) pressure p and the temperature T of the fluid which are two-dimensional now are given as

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{4}$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial p}{\partial x} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - Mn(T_c - T)H\frac{\partial H}{\partial x} \tag{5}$$

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = \frac{\partial p}{\partial y} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} - Mn(T_c - T)H\frac{\partial H}{\partial y} - \frac{Ra}{Pr}T \tag{6}$$

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = P_z + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} \tag{7}$$

$$\begin{aligned}\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} &= Pr(u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}) - MnEcPr(\epsilon + T)H(u\frac{\partial H}{\partial x} + v\frac{\partial H}{\partial y}) \\ &\quad - EcPr(2\left(\frac{\partial u}{\partial x}\right)^2 + 2\left(\frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2)\ .\end{aligned} \tag{8}$$

The pressure equation can also be obtained by differentiating the $x-$ and $y-$components of the momentum equations and adding them with the use of continuity equation

$$\begin{aligned}\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} &= \frac{Ra}{Pr}\frac{\partial T}{\partial y} - \left(\frac{\partial u}{\partial x}\right)^2 - \left(\frac{\partial v}{\partial y}\right)^2 - 2\frac{\partial v}{\partial x}\frac{\partial u}{\partial y} - Mn(\frac{\partial T}{\partial x}H\frac{\partial H}{\partial x} + \frac{\partial T}{\partial y}H\frac{\partial H}{\partial y}) \\ &\quad + Mn(T_c - T)(\left(\frac{\partial H}{\partial x}\right)^2 + \left(\frac{\partial H}{\partial y}\right)^2 + H\nabla^2 H)\ .\end{aligned} \tag{9}$$

The second term in the energy equation is the heating due to magnetization and the last term is the viscous dissipation which is a heat source caused by the friction between the fluid particles.

The terms $Mn(T_c - T)H\frac{\partial H}{\partial x}$ and $Mn(T_c - T)H\frac{\partial H}{\partial y}$ in the momentum equations are the components of the magnetization force so called Kelvin force. T_c is the Curie temperature of iron.

The last term in the $y-$ component of the momentum equation is the buoyancy force caused

 ISBN: 978-84-608-6082-2

by the gravitational force and wall temperatures difference.
The non-dimensional parameters entering the problem are

$$Mn = \frac{\mu_0 \chi H_0^2 (T_{hot} - T_{cold}) h^2}{\nu^2 \rho} \text{ (Magnetic number)}, \quad Pr = \frac{\rho c_p \nu}{k} \text{ (Prandtl number)}, \quad (10)$$

$$Ra = \frac{g \rho c_p \beta (T_{hot} - T_{cold}) h^3}{\nu k} \text{ (Rayleigh Number)}, \quad \epsilon = \frac{T_{cold}}{T_{hot} - T_{cold}} \text{ (Temperature Number)}, \quad (11)$$

$$Ec = \frac{\nu^2}{h^2 c_p (T_{hot} - T_{cold})} \text{ (Eckert Number)} \quad (12)$$

where, ν is the kinematic viscosity, ρ is the density, c_p is the specific heat, k is the thermal conductivity, χ is the magnetic susceptibility , β is the thermal expansion coefficient of the fluid; h is the width of the cavity, μ_0 is the magnetic permeability of vacuum ($\mu_0 = 4\pi \times 10^{-7}$Tm/A) and g is the magnitude of the gravitational acceleration ($g = -9.81\text{m}/s^2$). T_{hot} and T_{cold} are the temperatures of the hot and cold walls, respectively.
Magnetic number expresses the ratio of the magnetic forces and the inertia forces. Prandtl number is the ratio of the momentum and thermal diffusivities. Rayleigh number is the product of the Grashof number and the Prandtl number where Grashof number denotes the ratio of the buoyancy forces to viscous forces. Eckert number defines the kinetic energy of the flow relative to the boundary layer enthalpy difference.
To be able to see the flow patterns on the cross-section of the pipe we define stream function satisfying the continuity equation as $u = \frac{\partial \Psi}{\partial y}, \quad v = -\frac{\partial \Psi}{\partial x}$. Then, the stream function equation is given by

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} = \frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} \, . \quad (13)$$

Boundary conditions for the velocities are the no-slip wall conditions and the walls are natural streamlines. The side walls are heated from the left and cooled from the right. We approximate pressure boundary conditions through $x-$ and $y-$components of the momentum equations using forward difference for the pressure gradients which includes inner values and the DRBEM coordinate matrix $\mathbf{F}$ for all the other terms. The problem geometry and the boundary conditions are presented in Figure 1.

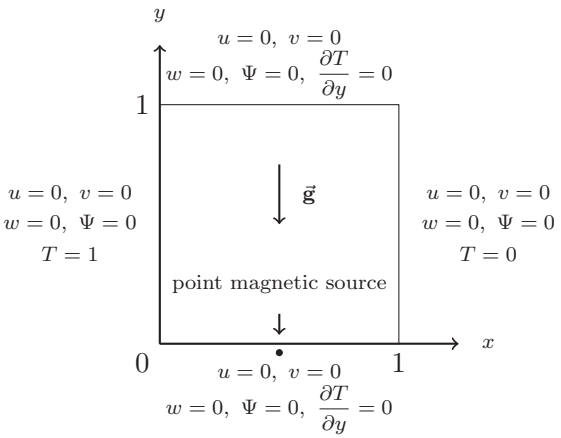


Figure 1: The problem geometry and the boundary equations

3 DRBEM Application

The dual reciprocity boundary element technique transforms Eqs. (5)-(9) and (13) into the boundary integral equations using the fundamental solution of Laplace equation ($u^* = (1/2\pi)ln(1/r)$) [6]. Taking all the terms other than Laplacian as inhomogeneity, weighting the equations with the fundamental solution u^* and using the Divergence theorem two times we obtain

$$c_i R_i + \int_\Gamma R q^* d\Gamma - \int_\Gamma u^* \frac{\partial R}{\partial n} d\Gamma = - \int_\Omega b_R u^* d\Omega \tag{14}$$

where R denotes u, v, w, T, p or Ψ and b_R is the right hand-side of each corresponding equation for R. $\Gamma = \partial\Omega$ and $q^* = \partial u^*/\partial n$. The value of c_i at the source point i is $c_i = \theta/2\pi$ where θ is the internal angle at the point i in radians.
We approximate b_R by the radial basis function $f_j(r) = 1 + r_j$ which is connected to the particular solution as $\nabla^2 \hat{u}_j = f_j$, [6] and r_j is the distance between the source and field points. Then, we get

$$b_R = \sum_{j=1}^{N+L} (\alpha_R)_j \nabla^2 \hat{u}_j \ . \tag{15}$$

Here $(\alpha_R)_j$'s are the undetermined coefficients for the approximation of the right hand-side b_R. N and L are the number of boundary and the interior nodes, respectively.
Applying the Divergence theorem also to the right hand-side of Eq. (14) and discretizing

the boundary with constant elements we obtain

$$c_i R_i + \sum_{k=1}^{N} \int_{\Gamma_k} R q^* d\Gamma - \sum_{k=1}^{N} \int_{\Gamma_k} u^* \frac{\partial R}{\partial n} d\Gamma = \sum_{j=1}^{N+L} (\alpha_R)_j (c_i \hat{u}_{ij} + \sum_{k=1}^{N} \int_{\Gamma_k} \hat{u}_j q^* d\Gamma - \sum_{k=1}^{N} \int_{\Gamma_k} u^* \frac{\partial \hat{u}_j}{\partial n} d\Gamma) \tag{16}$$

Eq. (15) is used to determine the unknown coefficients in terms of the DRBEM coordinate matrix $\mathbf{F}$ which is constructed by taking f_j's as columns.

$$\alpha_R = \mathbf{F}^{-1} b_R \tag{17}$$

Writing the Eq (16) for all boundary (with $c_i = 1/2$) and L interior nodes (with $c_i = 1$) and using (17) we arrive at DRBEM discretized matrix-vector equations

$$\mathbf{H}u - \mathbf{G}\frac{\partial u}{\partial n} = (\mathbf{H}\hat{\mathbf{U}} - \mathbf{G}\hat{\mathbf{Q}})\mathbf{F}^{-1}\{\frac{\partial p}{\partial x} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - Mn(Tc - T)H\frac{\partial H}{\partial x}\} \tag{18}$$

$$\mathbf{H}v - \mathbf{G}\frac{\partial v}{\partial n} = (\mathbf{H}\hat{\mathbf{U}} - \mathbf{G}\hat{\mathbf{Q}})\mathbf{F}^{-1}\{\frac{\partial p}{\partial y} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} - Mn(T_c - T)H\frac{\partial H}{\partial y} - \frac{Ra}{Pr}T\} \tag{19}$$

$$\mathbf{H}w - \mathbf{G}\frac{\partial w}{\partial n} = (\mathbf{H}\hat{\mathbf{U}} - \mathbf{G}\hat{\mathbf{Q}})\mathbf{F}^{-1}\{P_z + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y}\} \tag{20}$$

$$\begin{aligned}\mathbf{H}T - \mathbf{G}\frac{\partial T}{\partial n} = (\mathbf{H}\hat{\mathbf{U}} - \mathbf{G}\hat{\mathbf{Q}})\mathbf{F}^{-1}\{&Pr(u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}) - MnEcPr(\epsilon + T)H(u\frac{\partial H}{\partial x} + v\frac{\partial H}{\partial y}) \\ &- EcPr(2(\frac{\partial u}{\partial x})^2 + 2(\frac{\partial v}{\partial y})^2 + (\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y})^2)\}\end{aligned} \tag{21}$$

$$\begin{aligned}\mathbf{H}p - \mathbf{G}\frac{\partial p}{\partial n} = (\mathbf{H}\hat{\mathbf{U}} - \mathbf{G}\hat{\mathbf{Q}})\mathbf{F}^{-1}\{&\frac{Ra}{Pr}\frac{\partial T}{\partial y} - (\frac{\partial u}{\partial x})^2 - (\frac{\partial v}{\partial y})^2 - 2\frac{\partial v}{\partial x}\frac{\partial u}{\partial y} \\ &- Mn(\frac{\partial T}{\partial x}H\frac{\partial H}{\partial x} + \frac{\partial T}{\partial y}H\frac{\partial H}{\partial y}) + Mn(T_c - T)((\frac{\partial H}{\partial x})^2 + (\frac{\partial H}{\partial y})^2 + H\nabla^2 H)\}\end{aligned} \tag{22}$$

$$\mathbf{H}\Psi - \mathbf{G}\frac{\partial \Psi}{\partial n} = (\mathbf{H}\hat{\mathbf{U}} - \mathbf{G}\hat{\mathbf{Q}})\mathbf{F}^{-1}\{\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\}\ . \tag{23}$$

Here,

$$\mathbf{H}_{ij} = c_i\delta_{ij} + \int_{\Gamma_j} q^* d\Gamma_j, \quad \mathbf{G}_{ij} = \int_{\Gamma_j} u^* d\Gamma_j, \quad \mathbf{G}_{ii} = \frac{l}{2\pi}(ln(\frac{2}{l}) + 1) \tag{24}$$

where δ_{ij} is the Kronecker delta and l is the length of the element. The matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{Q}}$ are constructed by taking each vector $\hat{u}_j$ and $\hat{q}_j = \partial\hat{u}_j/\partial n$ as columns, respectively.

All the space derivatives on the right hand-sides of Eqs. (18)-(23) are approximated by the coordinate matrix $\mathbf{F}$ as

$$\frac{\partial A}{\partial \eta} = \frac{\partial \mathbf{F}}{\partial \eta}\mathbf{F}^{-1}A, \qquad \frac{\partial^2 A}{\partial \xi \partial \eta} = \frac{\partial \mathbf{F}}{\partial \eta}\mathbf{F}^{-1}\frac{\partial \mathbf{F}}{\partial \xi}\mathbf{F}^{-1}A \tag{25}$$

with A being u, v, w, p or T , and ξ and η denote x or y. Discretized system of Eqs. (18)-(23) are solved iteratively by taking initial velocity and temperature of the fluid.

4 Numerical Results

Flow and thermal convection behaviors are studied for varying values of Magnetic number Mn and Rayleigh number Ra. Viscous dissipation effect is also shown. For our biomagnetic fluid model we set $Pr = 20$, $Ec = 1.25 \times 10^{-8}$, $\epsilon = 7$ [3]. The axial pressure gradient $Pz = -8000$ as in [2]. Magnetic source is placed at $(a, b) = (0.5, -0.05)$ below the cavity.
The proposed numerical scheme is validated for natural convection flow of air in a square cavity without magnetic effect. For this, we set $Pr = 0.7$ and $Ra = 10^3$ numerical results are found to be in good agreement with the results of Lo et al. [7]
Figures 2-3 display the flow characteristics for increasing Magnetic number; $Mn = 5, 200$ when there is no buoyancy force effect; $Ra = 0$. The numerical results reveal that as Mn increases the flow velocities and the pressure increase in magnitude at the same time the axial velocity shows a retardation around the magnetic point source. Streamlines and the $u-$ velocity consist of two symmetric vortices. Vertical velocity spreads through the cavity and as Mn increases new vortices appear close to the bottom corners of the cavity. Pressure highly concentrates around the point magnetic source. An increase in Mn shifts the isotherms through the hot and cold walls. For $Mn = 200$ they are divided almost symmetrically leaving the center of the cavity with an average temperature of the hot and cold wall temperatures.
When magnetic source is present ($Mn = 10$) and Rayleigh number increases from $Ra = 10^2$ to $Ra = 10^5$ (Figures 4-5) the flow, pressure and temperature behaviors completely change, especially for $Ra = 10^5$ only buoyancy force controls the flow. The symmetry in streamlines and $u-$velocity is destroyed enlarging one of the vortices because thermal convection dominates the magnetic effect. The right vortex center in $u-$velocity move upwards and it squeezes on the left upper corner. Isotherms shift to the hot wall and center of the cavity is cooled. This means that high thermal convection blocks the heat transfer from the hot wall. The $v-$velocity also concentrates to the hot wall. Pressure around the magnetic source moves to the top wall as Ra increases. The behavior of the axial velocity do not alter much.
For a larger magnetic number $Mn = 80$ and varying Rayleigh number $Ra = 10^2 - 10^5$ (Figures 6-7) magnetic source this time dominates the thermal convection up to $Ra = 10^5$.

 ISBN: 978-84-608-6082-2

The Rayleigh number effect on the temperature of the fluid is the formation of the thin layers close to the hot and cold walls of the cavity. When $Ra = 10^5$ again the symmetry in streamlines and the velocities disappears. The left vortex in streamlines shifts to the hot wall leaving its place to the right vortex. Pressure around the source looses its effect. Thermal convection is still observed near the cold wall.

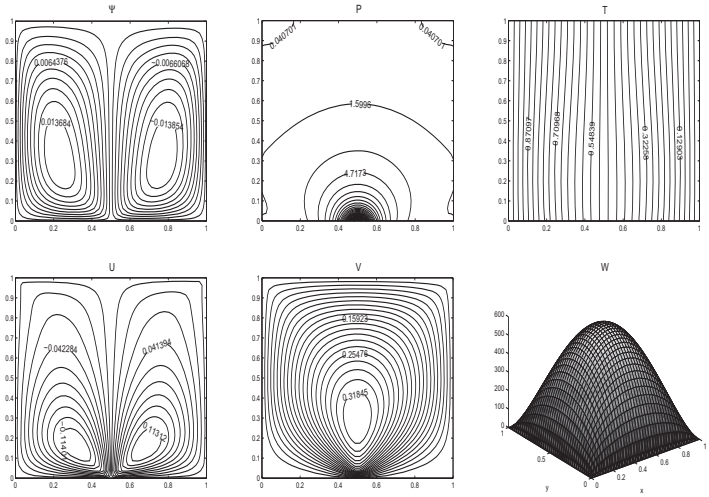


Figure 2: Viscous dissipation is neglected. $N = 160$, $Ra = 0$, $Mn = 5$

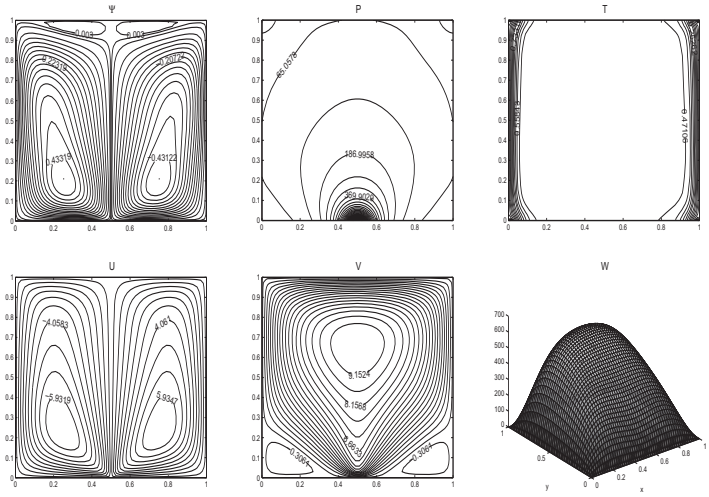


Figure 3: Viscous dissipation is neglected. $N = 200$, $Ra = 0$, $Mn = 200$

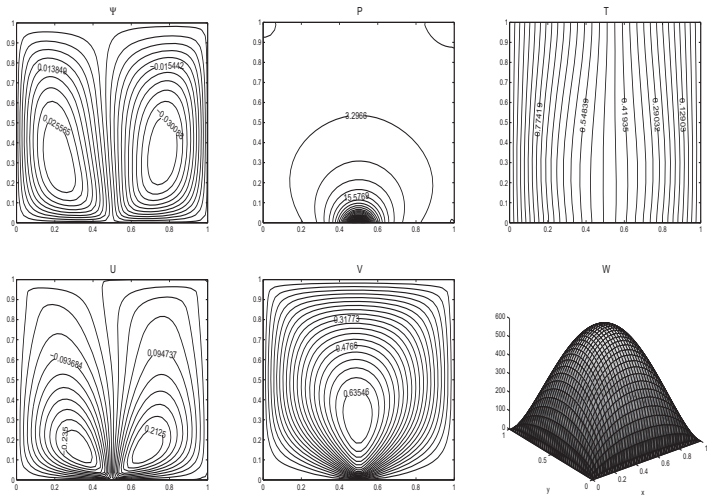


Figure 4: Viscous dissipation is neglected. $N = 160$, $Mn = 10$, $Ra = 10^2$

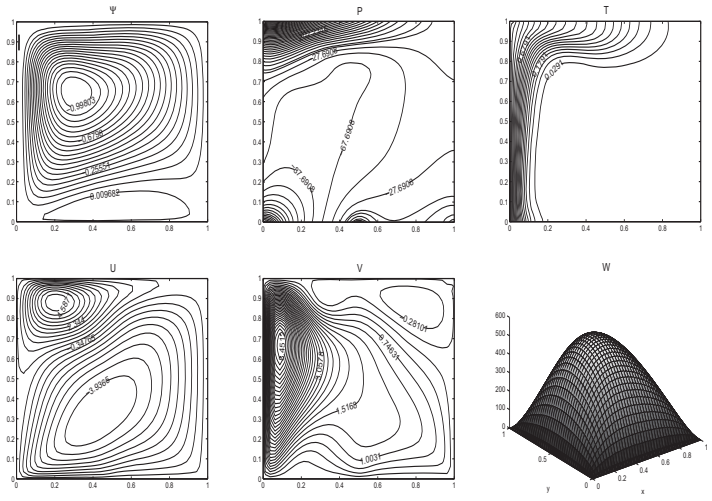


Figure 5: Viscous dissipation is neglected. $N = 160$, $Mn = 10$, $Ra = 10^5$

Finally, viscous dissipation effect on the flow, pressure and the temperature is visualized in Figures 8-9, respectively for $Mn = 10$ and 80 when $Ra = 10^5$. When magnetic source intensity is small ($Mn = 10$) flow and temperature behaviors do not change with viscous dissipation when it is compared with Figure 5. We observe the viscous dissipation effect on isotherms only when both magnetic and buoyancy forces are high ($Mn = 80$ and $Ra = 10^5$)

when it is compared with Figure 7. The convection from the right cold wall still extends through the center of the cavity. Magnitudes of the flow velocities and streamlines are reduced. As a result viscous dissipation retards the flow and the heat transfer in the cavity when the viscous fluid is dissipative. Also the $v-$velocity profile becomes nearly symmetric.

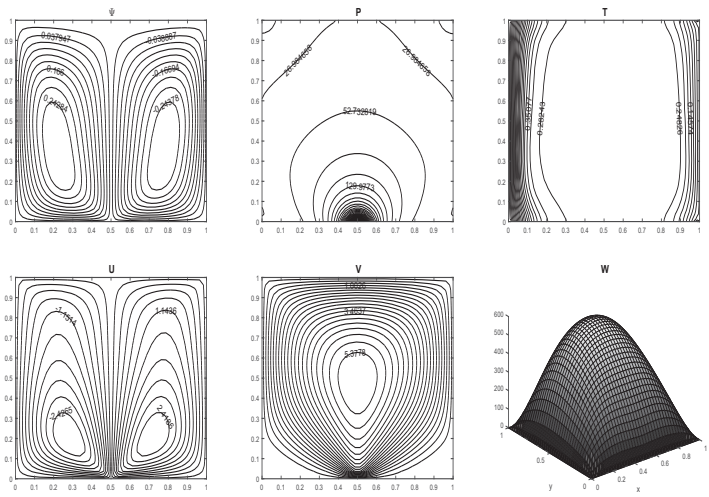


Figure 6: Viscous dissipation is neglected. $N = 160$, $Mn = 80$, $Ra = 10^2$

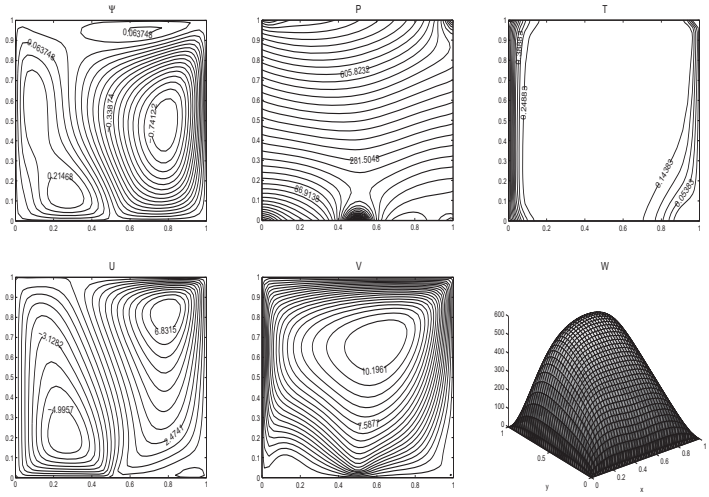


Figure 7: Viscous dissipation is neglected. $N = 160$, $Mn = 80$, $Ra = 10^5$

 ISBN: 978-84-608-6082-2

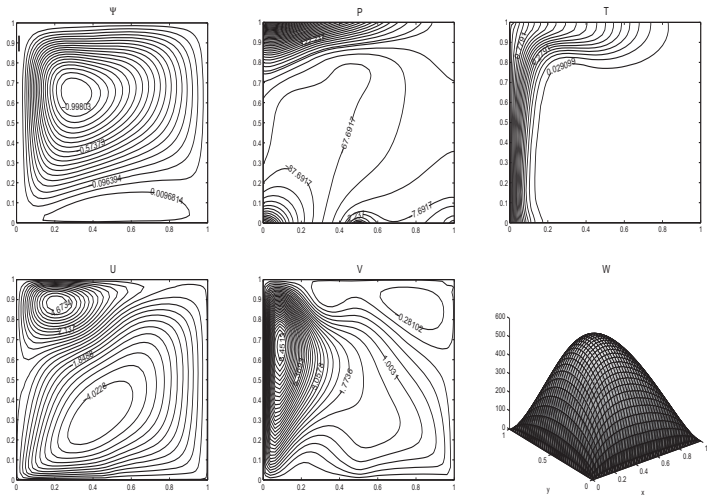


Figure 8: Viscous dissipation is included. $N = 160$, $Mn = 10$, $Ra = 10^5$

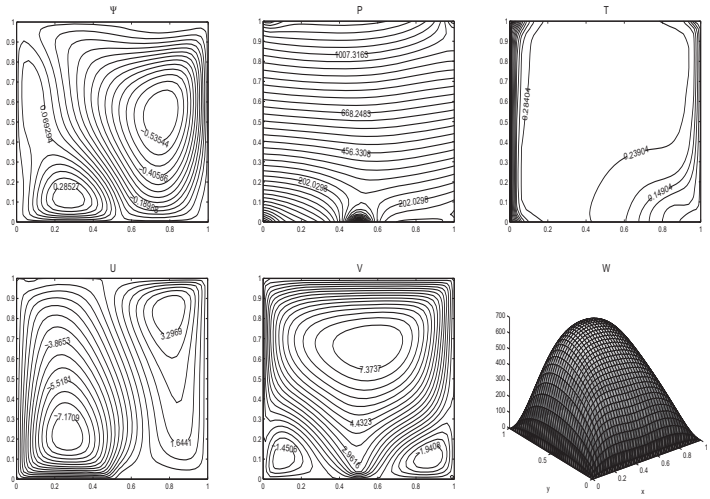


Figure 9: Viscous dissipation is included. $N = 160$, $Mn = 80$, $Ra = 10^5$

5 Conclusion

Forced convection of biomagnetic viscous fluid under the effect of a point source magnetic field is simulated by DRBEM. The numerical results are presented in terms of velocity,

pressure and the temperature of the fluid. The numerical results reveal that increasing magnetic field intensity divides the isotherms almost symmetrically through the hot and cold walls. When Rayleigh number is high the thermal convection beats the magnetic effect and change the flow characteristics significantly if the magnetic source is not strong enough. When Magnetic number is increased the symmetry in the flow is destroyed especially when Ra is large. The viscous dissipation effect is observed when both magnetic and buoyancy forces are high, the flow is retarded and the heat transfer is reduced.

References

[1] R. E. Rosensweig, *Ferrohydrodynamics*, Dover Publications, Mineola New York, 2014.

[2] E. E. Tzirtzilakis, V. D. Sakalis, N. G. Kafoussias and P. M. Hatzikonstantinou, *Biomagnetic fluid flow in a 3D rectangular duct*, International Journal for Numerical Methods in Fluids **44** (2004) 1279–1298.

[3] V. C. Loukopoulos and E. E. Tzirtzilakis, *Biomagnetic channel flow in spatially varying magnetic field* , International Journal of Engineering Science **42** (2004) 571–590.

[4] N. A. Idris, N. Amin, and H. Rahmat, *Effect of gravitational acceleration on unsteady biomagnetic fluid flow*, Applied and Computational Mathematics **3(6)** (2014) 285–294.

[5] C. A. J. Fletcher, *Computational Techniques for Fluid Dynamics 2*, Springer, Berlin, 1991.

[6] P. W. Partridge, C. A. Brebbia and L. C. Wrobel, *The Dual Reciprocity Boundary Element Method*, Computational Mechanics Publications, Southampton Boston, 1992.

[7] D. C. Lo, D. L. Young and C. C. Tsai, *High resolution of 2D natural convection in a cavity by the DQ method*, Journal of Computational and Applied Mathematics **203** (2007) 219–236.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Numerical simulation of cable truss systems using meshfree RBF method

Stanislav Simonenko[1], Jose Antonio Loya[1], Marcos Rodriguez Millan[1] and Philippe Angot[2]

[1] *Mecanica de Medios Continuos y Teoria de Estructuras, Universidad Carlos III de Madrid*

[2] *Institut des Mathmatiques de Marseille, Universite de Marseille*

emails: stan@bridgemaths.net, jloya@ing.uc3m.es, mrmillan@ing.uc3m.es, philippe.angot@univ-amu.fr

Abstract

In this paper are exposed algorithms and results, realized in the UC3M-lib modeling library, used for cable truss computations, suitable for hanging bridge engineering. The cable motion is simulated using meshfree RBF method, applied to hanging truss theory.

1 Introduction

The exact analysis of cable truss systems under a load function (continuous or not) has a great practical interest - overall, in hanging bridge engineering. The traditional finite elements methods, used in practice for hanging trusses of bridges, have following disadvantages:

- The section of cables, small, compared to characteristical dimensions of the structure, requires the use of big number of tiny elements, regardless of wether the cable system is analyzed with the whole structure, or separately. This follows to a great computational cost and difficult contact modelling between parts of hanging truss.
- The mesh generation for a hanging structure modeling requires a choice: uniform 3D elements, or a ficticious domain approach, making some elements bidimensional. This case modifies the initial problem, and is suitable only for

 ISBN: 978-84-608-6082-2

- Statically undefined structures (including many hanging trusses used in actual civil engineering) cannot be trated using finite elements methods, that is why FEM in hanging structures analysis has limited use for stayed and continuous load monocable structures.

Meshfree methods, applied to hanging bridge engineering, offer an alternative that avoids common problems. First, a structure is considered as a point cloud, with specific boundary conditions. Second, a immersed geometry approach allows to define any systems of loads on the cable system (that is usefull for practical cases of normalized analysis).

The mentionned algorithms are integrated into the UC3M-lib modeling library.

2 Meshfree RBF method for partial differential equations solving

A modelled geometry is considered is a set of nodes $X = \{\vec{x}_1, \ldots, \vec{x}_M\} \subset \Omega$ and $X_b = \{\vec{x}_{M+1}, \ldots, \vec{x}_N\} \subset \partial\Omega$, in a computational domain Ω. We consider $\mathcal{L}$ a spatial differential operator, $u(\vec{x})$ a function on Ω, with a PDE in the domain:

$$\begin{cases} \mathcal{L}u(\vec{x}) = f(\vec{x}) & \text{inside the domain } \Omega \\ \mathcal{B}u(\vec{x}) = g(\vec{x}) & \text{on the boundary } \partial\Omega \end{cases} \tag{1}$$

We consider $s(\vec{x})$ the approximation of the function $u(\vec{x})$ in the domain Ω, on the given set of nodes, constructed as a combination of some radial basis functions $\varphi_i(\vec{x})$, though

$$u(\vec{x}) \simeq s(\vec{x}) = \sum_{i=1}^{N} \alpha_i \varphi_i(\vec{x}) \tag{2}$$

where α_i is the respective weight of each basis function. The $\vec{\alpha}_i$ weights vector is defined by solving the matricial equation of colocation of known function values in centers u_D (that is, in our case, Dirichlet boundary conditions of simulation):

$$\vec{\alpha} = A^{-1}\vec{u}_D \tag{3}$$

where A is the colocation matrix. For radial basis function we may chose Multiquadric ($\varphi(\vec{x}, \vec{y}) = \sqrt{\|\vec{x}, \vec{y}\| + c^3}$), or, for the practical bridge engineering issues the most appropriate is the Power Spline ($\varphi(\vec{x}, \vec{y}) = -(\|\vec{x}, \vec{y}\|)^c$). In both cases the variable $c \in \mathbb{R}^+$ is called **shape parameter**. The shape parameter allows getting the approximation method accurate, and it's choice is important. A specific algorithm of it's choice for a cable truss is given in the respective part.

3 Elastodynamic problem: single cable

The general formulation of an elastodynamic problem is the following [1]. At a time iteration t_i the relation between the deformation and the stress in the system on the domain Ω is

$$\vec{\varepsilon} = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1 & \frac{\nu}{1-\nu} & \frac{\nu}{1-\nu} & 0 & 0 & 0 \\ \frac{\nu}{1-\nu} & 1 & \frac{\nu}{1-\nu} & 0 & 0 & 0 \\ \frac{\nu}{1-\nu} & \frac{\nu}{1-\nu} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1-2\nu}{2(1-\nu)} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1-2\nu}{2(1-\nu)} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1-2\nu}{2(1-\nu)} \end{bmatrix} \vec{\sigma} \tag{4}$$

and on the boundary $\partial\Omega$ the Dirichlet boundary conditions are $\vec{\varepsilon} = \vec{\varepsilon}_B$. For the resolution the spacial variables are separated, and each is approximated as:

$$\begin{cases} u(\vec{x}) = \sigma_{k=1}^{N} a_k \phi\left(\|\vec{x} - \vec{x}_k\|, c_u\right) \\ v(\vec{x}) = \sigma_{k=1}^{N} b_k \phi\left(\|\vec{x} - \vec{x}_k\|, c_v\right) \\ w(\vec{x}) = \sigma_{k=1}^{N} c_k \phi\left(\|\vec{x} - \vec{x}_k\|, c_w\right) \end{cases} \tag{5}$$

The equation 4 is solved locally on given stencils, in aim to optimize the shape parameter, by the following algorithm [2] on shape parameters of each dimension:

1. **Stencil check**: first, the stencil is considered as equal to domain.
2. **Data points check**: by loop on data points we make stencils of at least 3 of them. Neighborhood nodes are added to the stencil, to avoid the non-overlapping of the stencils on the domain.
3. **Error vectors construction** *The shape parameter $c = c_{x,y,z}$ is considered as a parsed variable.* The error vector of the stencil S_i with n_d nodes is computed as

$$\vec{E} = \begin{bmatrix} E_1, \\ \vdots \\ E_{n_d} \end{bmatrix} \tag{6}$$

   where $E_j = f_j - I_j(\vec{x_j})$, and I_j is the interpolation function on the reduced stencil $S_i \setminus (\vec{x}_j, f_j)$. Here is the necessity of having enough data points in the stencil. Thus, the error vector is the function of shape parameter $\vec{E}(c)$.
4. **Minimizing error on the stencil** by finding c_{min} to minimize $\|\vec{E}(c)\|$.

The particularity of a cable modeling are an imposed load and displacement on several nodes, and the Young's modulus as a function of displacement. Graphical results for a cable of a unitary length, released after a load on a moment 0, are shown as following:

 ISBN: 978-84-608-6082-2

(a) $t = 0.0$

(b) $t = 0.03$

(c) $t = 0.06$

(d) $t = 0.09$

(e) $t = 0.12$

(f) $t = 0.15$

Figure 1: Dynamic simulation using global method with GA RBF of a cable, submited to a force $P = 100$ that is released at $t = 0$. The cable starts oscillations.

4 Hanging truss modeling using fictitious domain approach

A hanging truss is a semi-solid system, using rods (cables) and nodes, connecting them. The number of unknown reactions is computed as

$$R = 2r_2 + 4r - 3 + 2(m-1)r_m \tag{7}$$

So, for a system composed of n rods, $3(n-1)$ static equations can be composed, containing R unknown reactions. In case when $R > 3(n-1)$ the system is statically undefined. In this case areas containing 3 rods are considered as fictitious domains and the problem is expanded to them.

Two algorithms are used in the numerical simulation:

- **Stencil selection**: the general loop on all nodes separates the cable rods, i.e. distances between nodes, and defines their characteristic diameter. The diameter is used to define contact conditions, if necessary.
- **Fictitious domain definition**: the 3-cables systems with at least 2 attach points may be considered as fictitious domain, with the boundary, delimited by these cables. It allows us the uniformization of properties, such as Young's modulus, and a computation of the sub-truss as of a solid.

For a point $\vec{x} \in \Omega$ in the computation domain was defined the biggest sphere radius, that do not contain nodes, laying in other cylinders, that one formed by $\vec{x}$ and the main direction of his neighbors. The main direction of his neighbors was defined by vector, where the concentration of neighbor nodes were maximal.

The numerical experiment was carried out in the following order: first, the load was placed in the point F of a simple X-truss, at time $t = 0$ it was released. Then 10 iterations with the step $\Delta t = 0.05$ were done. The results are presented on the figures 2a - 2d. In the case of parallel computing sensors we also placed near the F point for capturing displacement.

References

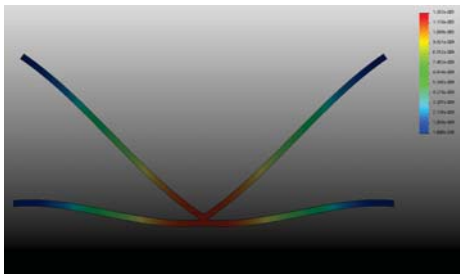
[1] S. Simonenko, V. Bayona, M. Kindelan , *Optimal shape parameter for the solution of elastostatic problems with the RBF method*, J. Eng. Math. **85(1)** (2014) 115-129

[2] S. Rippa, *An algorithm for selecting a good value for the parameter c in radial basis function interpolation*, Adv. Comp. Math. **11(2-3)** (1999) 193-210

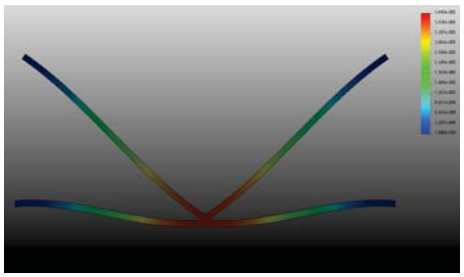
[3] V. K. Kachurin, A. V. Bragin, B. G. Erunov, *Hanging bridges design*, KTP Editions (2012)

[4] B. Adams, M. Ovsjanikov, M. Wand, *Meshless modeling of deformable shapes and their motion*, SCA 2008 Proceedings (2008)

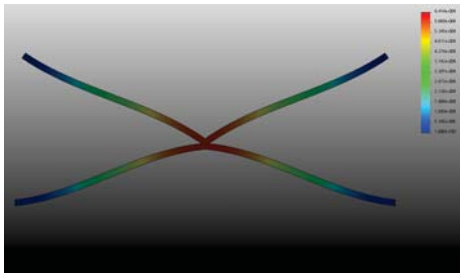
[5] L. Liao and B. Du, *Finite Element Analysis of Cable truss Structures*, 51st AIAA/ASME/ASCE/AHS/ASC Struct. and Mat. Conf. (2010)



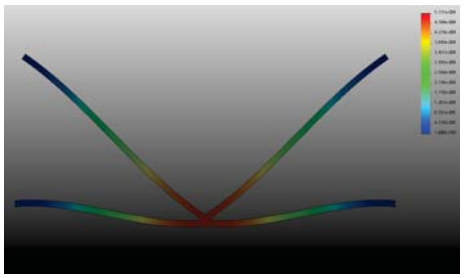
(a) $t = 0.0$



(b) $t = 0.05$



(c) $t = 0.15$



(d) $t = 0.25$

Figure 2: Results of dynamic parallel simulation using meshfree RBF method of a hanging truss, initially submitted to a force and released at $t = 0$. The displacements on this and following images witness a vertical oscillation, but without entering in second degree eigen modes.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Melting mercury with a quantum model: relativistic effects on thermodynamics

Krista G. Steenbergen[1], Elke Pahl[2], Florent Calvo[3] and Peter Schwerdtfeger[1]

[1] *Center of Theoretical Chemistry and Physics, The New Zealand Institute for Advanced Study, Massey University Auckland, New Zealand*

[2] *Centre for Theoretical Chemistry and Physics, Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand*

[3] *Laboratoire Interdisciplinaire de Physique, Université Grenoble Alpes and CNRS, St Martin d'Héres, France*

emails: kgsteen@gmail.com, E.Pahl@massey.ac.nz, florent.calvo@ujf-grenoble.fr, P.A.Schwerdtfeger@massey.ac.nz

Abstract

Using two distinct quantum simulation methods (Diatomics-in-Molecules and density functional theory molecular dynamics), we present the results of bulk and cluster melting simulations for mercury. We compare the relativistic and non-relativistic models, demonstrating that relativity plays a significant role in the thermodynamics of this interesting metal. We further compare the structure and dynamics of the solid and liquid phases of mercury at the phase transition temperature.

Key words: melting, density functional theory, molecular dynamics, Monte Carlo simulations, mercury, relativity

1 Simulating mercury melting

As the only elemental metal that exists in the liquid state at standard temperature and pressure, mercury's thermodynamic properties have long been of interest to the scientific community. It has been proposed that the origin of this phenomenon lies in mercury's relativistic nature. As a closed-shell atom (filled $5d^{10}$ shell), relativistic spin-orbit coupling

effects are minimal; however, scalar relativity leads to a contraction of mercury's $6s$ shell, which in turn stabilizes the $6s$ orbital while destabilizing the $5d^{10}$ shell. Our group has recently completed mercury melting simulations at ambient pressure using a Monte Carlo Diatomics-in-Molecules method (DIM).[1, 2] With the relativistic DIM model, the simulation yielded a melting temperature of 250 K, in good agreement with the experimental value of 234 K.[3] Remarkably, the non-relativistic model melts at 355 K – more than 105 K greater than the relativistic model.[3] This significant finding means that, without relativistic effects, mercury would be a solid at room temperature. We have also been able to extend this model to melting small mercury clusters where, even at finite size, relativity alters the cluster melting temperature by up to 176 K.[4]

While these results are ground-breaking, further testing has revealed that the DIM model may not be well-suited for increasing pressures (higher than ambient). Additionally, the Monte Carlo approach to the DIM simulations makes dynamic and structural analysis of the solid and liquid phases challenging or impossible. These analysis methods are integral to understanding how and why the melting phase transition changes under pressure.

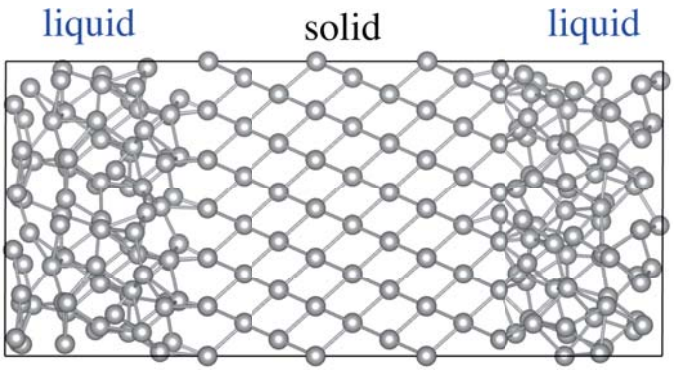


Figure 1: Interface pinning supercell for bulk mercury melting simulations, illustrating the two solid-liquid interfaces.

With relativity playing a large thermodynamic role, quantum simulations become requisite to accurately capture mercury's thermodynamic properties. Bulk melting simulations at the first-principles level of theory have only recently become computationally viable, as these calculations require a large periodic supercell (multiple unit cells), extensive simulation times, as well as some advanced simulation techniques to obtain realistic results. A new thermodynamic method, deemed "interface pinning", has recently made accurate simulations of bulk melting possible at the first-principles level of theory. This method

exploits the fact that the Gibbs free energy difference between the solid and liquid phase is zero at the melting phase transition. The solid and liquid phase are simulated together in the same simulation box which (due to periodic boundary conditions) gives two solid-liquid interfaces, as illustrated in Figure 1. As atoms change between the solid and liquid phases (according to which phase has the lowest Gibbs free energy), these two interfaces will move towards or away from one another. The solid-liquid interface is then "pinned" by an external field (a harmonic bias potential) which biases the system towards maintaining an equal number of solid and liquid atoms. The advancement of this simulation technique to the first-principles level of theory is ground-breaking and, for the first time, allows for highly accurate bulk melting simulations through DFT molecular dynamics (MD). This new technique has already been applied to Na, Mg, Al and Si, yielding excellent results for each of these materials.[5] Additionally, since the simulation utilizes the molecular dynamics algorithm, mercury's structural and dynamic response to temperature (such as δ_{rms}) can be analyzed.

In this talk, we present the bulk and cluster melting simulation results for both the relativistic and non-relativistic models. We compare and contrast Diatomic-in-Molecules simulation results with those obtained through DFT-MD interface pinning at ambient temperature. Additionally, we analyze dynamic measures, such as mean squared displacement (MSD) and δ_{rms}, in order to gauge the structural response to increasing temperatures and pressures.

Acknowledgements

This work has been supported by the Marsden Fund administered by the Royal Society of New Zealand, MAU1409.

References

[1] J. C. Tully, *J. Chem. Phys.* **1973**, *58*, 1396.

[2] H. Kitamura, *Eur. Phys. J. D* **2007**, *43*, 33–36.

[3] F. Calvo, E. Pahl, M. Wormit, P. Schwerdtfeger, *Angew. Chem. Int. Ed.* **2013**, *52*, 7583–7585.

[4] F. Calvo, E. Pahl, P. Schwerdtfeger, F. Spiegelman, *J. Chem. Theory Comput.* **2012**, *8*, 639–648.

[5] U. R. Pedersen, F. Hummel, G. Kresse, G. Kahl, C. Dellago, *Phys. Rev. B* **2013**, *88*, 094101.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Geometrical Interpretation of Complex Signals as a Tool to Study Fluctuations at Nanoscale

Bosiljka Tadić[1]

[1] *Department of Theoretical Physics, Jožef Stefan Institute*

emails: bosiljka.tadic@ijs.si

Abstract

By mapping the temporal sequence of a system's observable onto a mathematical graph, the objective analysis of the graph theory and algebraic topology techniques are applied to determine salient features of the underlying fluctuations phenomena. The methodology is implemented in nanoscale systems (i) to select the force-distance curves originating from the similar events in single-molecule force spectroscopy, moreover, (ii) to unravel the connection between the collective charge fluctuations and the enhanced conducting properties of nanoparticle assemblies within Coulomb blockade regime.

Key words: collective fluctuations, graph theory, algebraic topology of graphs
MSC 2000: 05C90, 37M10, 55M99

1 Introduction

Stochastic fluctuations of a physical observable, which can be recorded in the experiments or simulations, represents a breath of the complex system's dynamics [1]. In the nanoscale systems, the size of fluctuations accounts for a particularly important issue. Moreover, the underlying dynamical phenomena carry signatures of the collective effects that are responsible for the emergence of new properties of the assembly of nanoscale objects. Therefore, theoretical approaches to handling fluctuations and understanding their nature are of paramount importance in research of nanoscale systems. In this work, we extend the approach based on the mathematical concept of nanonetworks [5]. The approach consists of the mapping of the fluctuating signal onto a graph and utilizing the objective analysis within graph theory and the algebraic topology of graphs.

 ISBN: 978-84-608-6082-2

2 Mapping of complex signals to mathematical graphs

In recent years, exploring the time-series–networks duality has provided new insight into the behavior of complex dynamical systems [2, 3]. Different ways exist to map a sequence of events (data points of a fluctuating observable) onto mathematical graphs, depending on the objectives of the studied problem. Consequently, different types of graphs result and suitable methods of graph theory can be utilized. We demonstrate the approach in two practical problems in nanoscience: selecting the relevant force-distance curves in the single-molecule force spectroscopy, and analyzing collective fluctuations in nanoparticle assemblies.

2.1 Similarity of FD curves revealed by mesoscopic structure of the graph

The profile of force-distance curves in single-molecule force spectroscopy, cf. curves in Fig. 1, contains information about stretching and unbinding events occurring during one full contact–pull-away cycle. For complex molecules and different geometries of the set-up, one can encounter unspecific binding as well as the events resulting from the internal molecular degrees of freedom, dynamics of the linkers as well as binding to the surface. Recently, we considered a selection of FD curves originating from different types of experiments where such situations can occur [4, 5]. The matrix of the Pearson's correlation coefficients (above

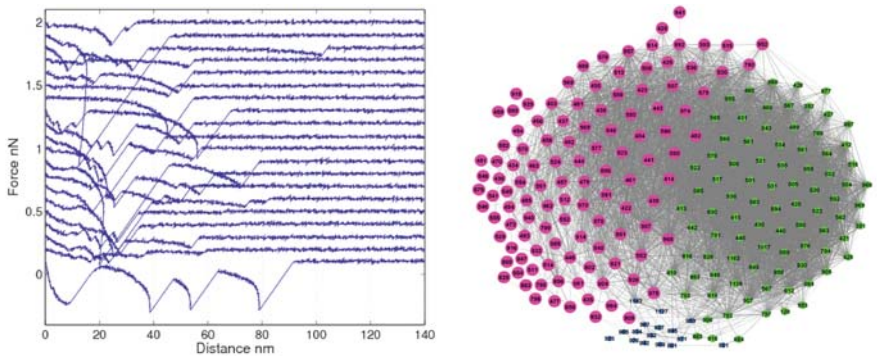


Figure 1: Examples of FD curves and the curves-correlation graph with weighted communities.

a threshold) can be regarded as an adjacency matrix of the graph [4]. A systematic analysis of the mesoscopic (community) structure of these graphs shows that the FD curves of a similar profile form a distinct group; such groups are detectable by the graph modularity optimization [7, 6]. An example of such structure is shown in Fig. 1. It appears that the curves group first according to the geometrical features of the set-ups and further by the type of binding site [4, 5]. We show that the curves originating from unspecific binding fall out of the group or line up with the events where the molecule is absent (binding to the Au-surface). The procedure then consists of collecting the ID's of the curves in the module

and selecting them from the experimental data; this set then provides reliable curves from which the binding force or bond kinetics can be determined. Interestingly, the standard graph measures [7] of the identified subgraphs are also different. For instance, the green (pale) and pink (dark) communities in Fig. 1 have the weighted degree 46.6 and 60.36, the average path length 1.489 and 1.784, and the graph density 0.57 and 0.38, respectively.

2.2 Collective charge fluctuations map onto graphs of higher topological complexity

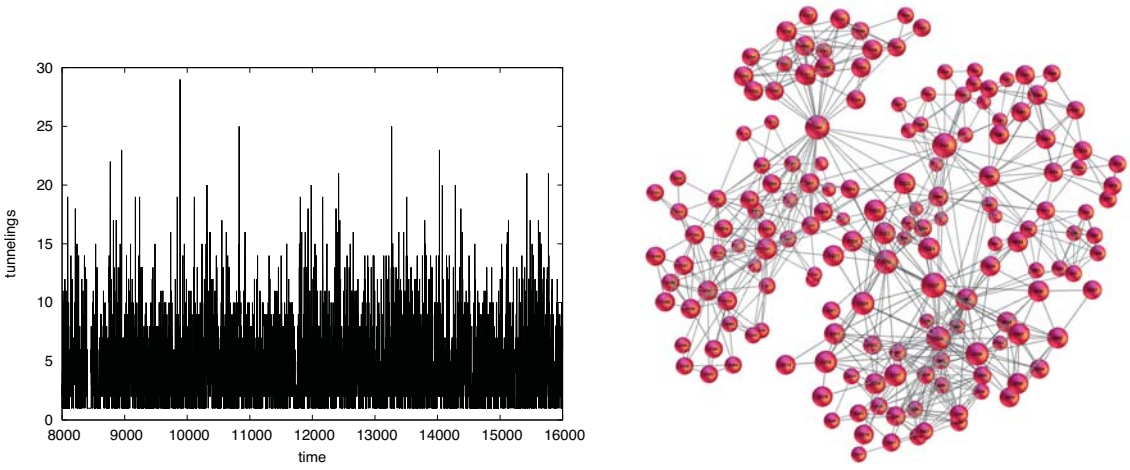


Figure 2: Charge fluctuations time series and a typical simplicial complex in the visibility graph.

In nanoparticle assemblies conducting via single-electron tunnelings within Coulomb blockade regime, the emergent properties which are manifested as large $I(V) \sim (V - V_T)^{\zeta}$ nonlinearity $\zeta \in (2, 4)$, are shown to be related to the architecture of the assembly [8, 9]. Recently, we have investigated the impact of a particular type of structural elements in the assembly that enhance the I-V nonlinearity [10]. The underlying mechanisms involve the enhanced collective fluctuations of charge. Using the 'visibility' method [2, 3], the time series of the number of tunnelings are mapped onto graphs, which represent the connections among the sequence of states in the phase space. Applying the methods of algebraic topology of graphs [11], we investigate the higher order structures (simplexes or cliques of high dimension) and their aggregates–simplicial complexes [12] in these graphs, cf. Fig. 2. The corresponding topological measures [10], characterize the phase space organization behind the collective charge fluctuations. We show that changes in the architecture of the assembly leading to higher I-V nonlinearity are compatible with stronger collective fluctuations, which are quantified by shifts in the topology measures of the phase-space graphs.

 ISBN: 978-84-608-6082-2

Acknowledgements

Work supported by the Program P1-0044 of the Research Agency of the Republic of Slovenia.

References

[1] A.N. Pavlov, V. S. Anishchenko, *Multifractal analysis of complex signals*, *Physics–Uspekhi* **50** (2007) 819–834.

[2] B. Luque, L. Lacasa, F. Ballesteros, A. Robledo, *Feigenbaum graphs: A complex network perspective of chaos*, *PLOS One* **6** (2011) e22411.

[3] M. Andjelković, N. Gupte, B. Tadić, *Hidden geometry of traffic jamming*, *Phys. Rev. E* **91** (2015) 052817.

[4] J. Živković, M. Mitrović, L. Janssen, H.Heuss, B. Tadić, S. Speller, *Network theory approach for data evaluation in the dynamic force spectroscopy of biomolecular interactions*, *Europhysics Letters* **89** (2010) 68004.

[5] J. Živković, B. Tadić, *Nanonetworks: The graph theory framework for modeling nanoscale systems*, *Nanoscale Systems MMTA* **2** (2013) 30–48.

[6] V.D. Blondel, J-L. Guillaume, R. Lambiote, E. Lefebvre, *Fast unfolding of communities in large networks*, *J. Stat. Mech. Theor. Experiment* **10** (2008) P10008.

[7] S. Dorogovtsev, *Lectures on Complex Networks*, Oxford Univ. Press, New York, 2010.

[8] M. O. Blunt, M. Šuvakov, F. Pulizzi, C. P. Martin, E. Pauliac-Vaujour, A. Stannard, A.W. Rushforth, B. Tadić, P. Moriarty, *Charge transport in cellular nanoparticle networks: meandering through nanoscale mazes*, *Nano Letter* **7(4)** (2007) 855-859.

[9] M. Šuvakov, B. Tadić, *Modeling collective charge transport in nanoparticle assemblies*, *J. Phys. Condens. Matt.* **22** (2010) 163201.

[10] B. Tadić, M. Andjelković, M. Šuvakov, *The influence of architecture of nanoparticle networks on collective charge transport revealed by the fractal time series and topology of phase space manifolds*, *J. Coupled Sys. Multisc.Dyn.* (2016) in press.

[11] J. Jonsson, *Simplicial Complexes of Graphs*, Lecture Notes in Mathematics, Springer-Verlag, Berlin, 2008.

[12] H.-J. Bandelt, V. Chepoi, *Metric graph theory and geometry: a survey,Contemporary Mathematics* **453** (2008) 49–86.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Taking out even more features from the input subset based on feature ranking

Antonio J. Tallón-Ballesteros[1,2] and Luís Correia[1]

[1] *Department of Computer Science, University of Lisbon. Campo Grande. Lisbon, 1749-016 Portugal*

[2] *Department of Languages and Computer Systems, University of Seville, 41012 Spain*

emails: antoniotallon@ciencias.ul.pt, Luis.Correia@ciencias.ulisboa.pt

Abstract

This paper introduces the application of an additional data preparation task via a feature ranking with an uncertainty measure to an initial reduced set that has been undergone to an initial feature subset selection with a correlation-based feature selection procedure guided by a stochastic algorithm. The proposal has been assessed in high-dimensional problems with a thousands of attributes. The results showed that NBTree (Nave Bayes Tree) is able to get better test accuracy results with a lower number of features, that is, the reduction in the input subspace is between a 25% and a 32%.

Key words: feature selection, classification, data mining, high-dimensionality, hybrid feature selection

1 Introduction

Machine learning is the field of scientific study that concentrates on induction algorithms and on other algorithms that can be said to learn [1]. Nowadays, the amount of information to be processed is unlimitedly increasing. Data preparation is a fundamental stage of data analysis [5]. We focus on high-dimensional supervised machine learning problems and hence one crucial task is feature selection. Basically, there are two ways to reduce the feature space depending on the use or not of a classifiers; these methods are called wrappers and filters, respectively. On the other hand, the output of the feature selection approach might be a subset of attributes or a ranking containing a weight for every feature; in this case the strategies are named feature subset selection (FSS) and feature ranking (FR).

This paper aims at analysing the convenience to diminish a feature subset with a feature reduction procedure based on an uncertainty measure in the context of high-dimensionality classification data sets.

2 Methodology and proposal

Problems with thousands of features could be reduced firstly via FSS or FR. The problem of FR is that the selection of individual features may not operate very well collectively and the number of attributes is also a handicap. On the other hand, FSS are able to get a subset that exhibits a good performance. Typically, the research has been conducted using FR and next FSS. Our methodology is based on the application of FSS and then the FS is dealt with. Recently, we have experienced with a stochastic FSS based on a scatter search procedure [4]. According to the results, correlation-based measures showed a better behaviour. We take as initial data sets the reduced data sets and we try to reduce them even more and at the same time to overcame the performance. The preliminary work was tested with five different seeds and therefore five pairs of training and test subsets were reached. The percentage of selected attributes was very small compared with the initial data sets containing several thousands of features. Now the idea is to diminish a bit the input subset space via FR. Several measures could be used in FR strategies and among them we have chosen symmetrical uncertainty. This metric has been widely applied in FSS but the same case has not happened in the context of FR.

3 Experimentation

Table 1 depicts the problems employed. The test bed consists of two multi-class high-dimensional data sets with thousands of features. The number of samples is around one and three hundred as usual in the Bioinformatics domain. The last ingredient is the number of classes that is between four and seven. Generally speaking, these problems are challenging due to their complexity and difficulty to get good performances. Last column specifies the average number of selected attributes with a correlation-based feature subset selection guided by a scatter search procedure (SS-CFS) using five different seeds.

SALL [2] problem is concerned with leukemia. SRBCT [3] represents small, round blue cell tumors (SRBCT) of childhood.

The starting point of this research is on the reduced problems with an initial FSS with SS-CFS. Now, we conduct a FR procedure via symmetric uncertainty measure to try to diminish the number of attributes. FR methods require a threshold to pick up only those attributes that meet a condition, that is, greater than, equal or lower than. Determining the threshold is a crucial task. According to the literature, there is not a common ground to do that in the sense that is unclear if the threshold should be defined for every data

Table 1: High-dimensional suppervised machine learning data sets

Data set	*Instances*	*Features*	*Classes*	$Features_{SS-CFS}$
SALL	327	12558	7	216.40
SRBCT	83	2308	4	107.60

Table 2: Initial and refined input space

Data set	*Thresholds for FR*	*Feature space size*		
		$SS-CFS$	$SS-CFS$ *and FR(Uncertainty)*	
			Threshold 1	*Threshold* 2
SALL	0.20, 0.25	216.40	167.60	152.60
SRBCT	0.40, 0.45	107.60	76.60	70.20
Averages				
Features		162.00	122.10	111.40
Reduction			25.68	32.12

set or classifier or for both simultaneously. We have done a preliminary experimentation with the first training set and four thresholds using the classifiers NBTree and PART. For SALL we have defined the following values: 0.15, 0.20, 0.25 and 0.3. On the other hand, SRBCT has been tested with 0.35, 0.40, 0.45 and 0.5. Finally, the chosen values were the pairs 0.20 and 0.25, and 0.40 and 0.45, respectively for the problems SALL and SRBCT. It is very important to underline that other values may be better but in this initial study we have limited the experimentation to the aforementioned values. Table 2 summarises the feature space concerning the initial situation and the final feature space after the additional data preparation procedure introduced in the current paper. The number of features has been reduced in average between a 25% and around a 32%. The reduction percentages are very interesting but we need to assess the new data sets with the classifiers that have been referenced before. It is very outstanding to remark that FR is applied only to the training set and the list of selected features is then project into the test set for every problem and seed.

4 Results and conclusions

Table 3 shows the test accuracy results in high-dimensionality classification data sets that have been get the current proposal along with the results using the initial data sets, that is, the full training and test sets have been preprocessed via a FSS method called SS-CFS. The results have been averaged with five different see due to the stochastic nature of the scatter search procedure included in SS-CFS. Every problem has been evaluated with a pair of thresholds. The last part of the table is devoted to shed light on the convenience of the

 ISBN: 978-84-608-6082-2

proposal and the recommendation for further works.

Firstly, SALL problem is improved with the classifier NBTree with any of the reported thresholds. Nevertheless, the second threshold only gets a feature set with a lower number of attributes. On the other hand, PART classifier reaches slightly better or worse results depending on the use of the second or the first one threshold, respectively. Secondly, SRBCT is a clear situation that in most cases the results are identical to the initial point but with a lower number of attributes. Next, a global overview is described. In average, for NBTree every threshold helps to find better results whereas for PART only with the second threshold minor enhancements take place. According to absolute numbers, in 3 out of 4 cases the new proposal is able to reach better average results. Taking a look to every data set and threshold, NBTree with the first threshold got a win and a tie; concerning the second threshold, NBTree and PART classifiers got better results twice.

Lastly, we can conclude that the new proposal is very suitable for NBTree because the test accuracy results are better in the majority of the cases and the feature space is reduced from 162 to the range 111-122. On contrary, PART experimented in some cases small improvements which means that PART is not very sensitive to features with a low performance according to symmetric uncertainty but it is a good idea for the future to start with a more reduced subset. To sum up, as first step a FSS such as SS-CFS has been used to get an initial reduced traini

As further work, we plan to assess the proposal with more supervised machine learning algorithms based also in trees or rules and with a higher number of problems, particularly with a similar complexity to the data sets that the current paper evaluated.

References

[1] R. Kohavi and F. Provost, *Glossary of terms*, Mach. Learn., **30(2-3)** (1998) 271–274.

[2] Eng-Juh Yeoh, Mary E Ross, Sheila A Shurtleff, W Kent Williams, Divyen Patel, Rami Mahfouz, Fred G Behm, Susana C Raimondi, Mary V Relling, Anami Patel, et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling*, Cancer cell, **1(2)** (2002) 133–143.

[3] Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson, et al. , *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*, Nature medicine, **7(6)** (2001)673–679.

 ISBN: 978-84-608-6082-2

Table 3: Test accuracy results in high-dimensionality supervised machine learning data sets

Data set	*Parameter values*	*Accuracy*			
		Classifier			
		NBTree		*PART*	
		Feature selection			
		$SS-CFS$	$SS-CFS$ *and* $FR(Unc.)$	$SS-CFS$	$SS-CFS$ *and* $FR(Unc.)$
SALL		77.56		79.76	
	Th. $1 = 0.20$		82.93		79.02
	Th. $2 = 0.25$		77.80		80.24
SRBCT		90.00		64.00	
	Th. $1 = 0.40$		90.00		64.00
	Th. $2 = 0.45$		91.00		64.00
Average		83.78		71.88	
Accuracy	*Th.* 1		86.47		71.51
	Th. 2		84.40		72.12
Improvements					
In average	*Th.* 1		*Yes*		*No*
	Th. 2		*Yes*		*Yes*
In punctual	*Th.* 1		*Once (other tie)*		*None(one tie)*
results	*Th.* 2		*Twice*		*Twice*

 ISBN: 978-84-608-6082-2

[4] A. J. Tallón-Ballesteros and A. Ibiza-Granados, *Simplifying pattern recognition problems via a scatter search algorithm* , International Journal for Computational Methods in Engineering Science and Mechanics, (in press) (2016).

[5] C. C. Aggarwal, *Data preparation. In C. C. Aggarwal, editor,* Data Mining, pages 27–62, Springer, Cham, 2015.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

On Cyclic Codes over $\mathbb{Z}_q + u\mathbb{Z}_q$

Fatih Temiz[1] **and Irfan Siap**[1]

[1] *Department of Mathematics, Faculty of Art and Sciences, Yıldız Technical University*

emails: ftemiz@yildiz.edu.tr, isiap@yildiz.edu.tr

Abstract

In this work, we first determine the ideal structure of the ring $\mathbb{Z}_q + u\mathbb{Z}_q$ where $q = 2^s$, s is any positive integer and $u^2 = 0$. Next, we give a formula that enumerates the number of ideals of this ring. Afterwards, we investigate the cyclic codes of odd length n over $\mathbb{Z}_q + u\mathbb{Z}_q$ and number of them by considering the ideal structure of the ring $(\mathbb{Z}_q + u\mathbb{Z}_q)[x]/\langle x^n - 1\rangle$ for $q = 2^2$ and $q = 2^3$.

Key words: cyclic codes, codes over rings
MSC 2000: MSC 94B15, MSC 94B05

1 Introduction

Cyclic codes are a significant class of linear codes due to their applicability to communication schemes. Since cyclic codes were introduced first by Prange in 1957 [4], they have been very interested to algebraic coding theorists. Hamming codes, BCH codes and Reed-Solomon codes are well-known codes over finite fields. Non-linear codes are more difficult to determine and process than linear codes in spite of the fact that they can be more effective. In 1994, Hammons et al. showed the relation between linear codes over the ring $\mathbb{Z}_4$ and some more acceptable non-linear binary codes [9]. Since the ring $\mathbb{F}_2 + u\mathbb{F}_2$ shared some good properties of $\mathbb{Z}_4$ and $\mathbb{F}_4$ it was considered to study cyclic codes in [1]. Thenceforth, cyclic codes over the various rings, some extensions of $\mathbb{F}_q$ and $\mathbb{Z}_2$ for instance, have been studied [8, 10, 7]. Recently, cyclic codes over $\mathbb{Z}_4 + u\mathbb{Z}_4$ and their $\mathbb{Z}_4$ images have been considered in [2]. Also minimum generating sets of cyclic codes over $\mathbb{Z}_q + u\mathbb{Z}_q$ were considered in [5] where q is a prime power and $u^2 = 0$. However, these studies do not consider the ideal structure of $\mathbb{Z}_q + u\mathbb{Z}_q$ or $(\mathbb{Z}_q + u\mathbb{Z}_q)[x]/\langle x^n - 1\rangle$. In this work, we consider the ring $\mathbb{Z}_q + u\mathbb{Z}_q$ where q is a power of 2 and $u^2 = 0$. We first determine the ideal structure of this ring. Next, we

give a formula that enumerates the number of ideals. Afterwards, we investigate the cyclic codes of odd length n and number of them and also size of them by considering the ideal structure of the ring $(\mathbb{Z}_q + u\mathbb{Z}_q)[x]/\langle x^n - 1\rangle$ for $q = 2^2$ and $q = 2^3$.

2 Preliminaries

Let q be 2^s for some positive integer s. $\mathbb{Z}_q + u\mathbb{Z}_q$ is the ring of elements $\{a + bu \,|\, a, b \in \mathbb{Z}_q\}$ where $u^2 = 0$, which is isomorphic to $\mathbb{Z}_q[u]/\langle u^2\rangle$. It is a commutative ring and it has characteristic q. The unit elements in this ring are precisely those elements $a + bu$ where a is a unit in $\mathbb{Z}_q$ [3]. We give the following theorem in order to determine all the ideals in $\mathbb{Z}_q + u\mathbb{Z}_q$.

Theorem 1 *The ideals of $\mathbb{Z}_q + u\mathbb{Z}_q$ are of the form:*

1. $\langle 2^i\rangle$ *for* $0 \le i \le s$
2. $\langle 2^i u\rangle$ *for* $0 \le i < s$
3. $\langle v2^i + 2^j u\rangle$ *for* $0 \le j < i < s$, $\gcd(2, v) = 1$ *and* $1 \le v < 2^{\min(i-j, s-i)}$
4. $\langle 2^i, 2^j u\rangle$ *for* $0 \le j < i < s$
5. $\langle 2^i + 2^j u, (v-1)2^i\rangle$ *for* $0 \le j < i < s$, $\gcd(2, v) = 1$ *and* $1 < v < 2^{\min(i-j, s-i)}$.

It can be easily seen that, $\mathbb{Z}_q + u\mathbb{Z}_q$ is not a principal ideal ring. The ideal $\langle 2, u\rangle$ cannot be generated by a single element for instance. Further $\langle 2, u\rangle$ is the unique maximal ideal i.e., $\mathbb{Z}_q + u\mathbb{Z}_q$ is a local ring. Also it is a non-chain ring since $\langle 2\rangle$ and $\langle u\rangle$ do not contain each other. Next, we give the number of ideal in this ring.

Theorem 2 *The number of all the ideals in $\mathbb{Z}_{2^s} + u\mathbb{Z}_{2^s}$ can be found via a recurrence relation*

$$K_{s+1} = 2K_{s-1} + 2s + 1$$

where $K_1 = 1$ and $K_2 = 3$ which yields

$$\sum_{i=1}^{s+1} i2^{\lfloor (s+1-i)/2 \rfloor}$$

We note that this recurrence relation also gives the number of nonempty subsequences $\{s(k)\}$ of $1, \ldots, n$ such that the difference sequence is palindromic and there is an explicit formula giving the result in [6].

Example 3 *Let us write down all the 37 ideals of $\mathbb{Z}_{32} + u\mathbb{Z}_{32}$:*

 ISBN: 978-84-608-6082-2

- $\langle 0\rangle, \langle 1\rangle, \langle 2\rangle, \langle 4\rangle, \langle 8\rangle, \langle 16\rangle$
- $\langle u\rangle, \langle 2u\rangle, \langle 4u\rangle, \langle 8u\rangle, \langle 16u\rangle$
- $\langle 2+u\rangle, \langle 4+u\rangle, \langle 12+u\rangle, \langle 4+2u\rangle, \langle 8+u\rangle, \langle 8+2u\rangle, \langle 24+2u\rangle, \langle 24+u\rangle, \langle 8+4u\rangle, \langle 16+u\rangle, \langle 16+2u\rangle, \langle 16+4u\rangle, \langle 16+8u\rangle$
- $\langle 2, u\rangle, \langle 4, u\rangle, \langle 4, 2u\rangle, \langle 8, u\rangle, \langle 8, 2u\rangle, \langle 8, 4u\rangle, \langle 16, u\rangle, \langle 16, 2u\rangle, \langle 16, 4u\rangle, \langle 16, 8u\rangle$
- $\langle 4+u, 8\rangle, \langle 8+u, 16\rangle, \langle 8+2u, 16\rangle$

3 Cyclic Codes over $\mathbb{Z}_q + u\mathbb{Z}_q$

Definition 4 *A linear code C of length n over $\mathbb{Z}_q + u\mathbb{Z}_q$ is a $(\mathbb{Z}_q + u\mathbb{Z}_q)-$submodule of $(\mathbb{Z}_q + u\mathbb{Z}_q)^n = \{(c_0, c_1, \ldots, c_{n-1}) \,|\, c_i \in \mathbb{Z}_q + u\mathbb{Z}_q$ for all $i = 0, 1, \ldots, n-1\}$.*

Definition 5 *A linear code C over $\mathbb{Z}_q + u\mathbb{Z}_q$ (we will denote by $\mathcal{R}$ from now on) is called a cyclic code if for each codeword $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1})$ the word $(c_{n-1}, c_0, \ldots, c_{n-2})$ obtained from $\mathbf{c}$ by cyclically shifting of coordinates is again in C i.e. the code is invariant under the cyclic shift operation.*

Conventionally, we match the codewords of a linear code over any ring R with polynomials in $R[x]$ bijectively as

$$(c_0, c_1, \ldots, c_{n-1}) \leftrightarrow c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}.$$

It is a well-known fact that, a linear code C of length n over a ring R is cyclic if and only if the corresponding polynomials of its codewords form an ideal structure in the ring $R[x]/\langle x^n - 1\rangle$. This motivates us to determine the ideal structure of the polynomial ring $\mathcal{R}[x]/\langle x^n - 1\rangle$ say $\Re_n$. Now, we give some definition and then give some lemmas that are analogue of lemmas which had been given to determine the ideal structure of $\mathbb{Z}_4[x]/\langle x^n - 1\rangle$ where n is odd, in [11]. Consider the map

$$\begin{aligned} - : \mathcal{R} &\to \mathbb{Z}_2 \\ a + bu &\mapsto \bar{a} = a \pmod 2 \end{aligned}$$

This map can be extended to a map between polynomial rings as follows:

$$\begin{aligned} \psi : \mathcal{R}[x] &\to \mathbb{Z}_2[x] \\ c_0 + c_1 x + \cdots + c_m x^m &\mapsto \bar{c}_0 + \bar{c}_1 x + \cdots + \bar{c}_m x^m. \end{aligned} \tag{1}$$

Definition 6 *The polynomials $f(x)$ and $g(x)$ are said to be coprime over $\mathcal{R}$ if there exist polynomials $a(x), b(x)$ over $\mathcal{R}$ such that $a(x)f(x) + b(x)g(x) = 1$.*

 ISBN: 978-84-608-6082-2

It is not difficult to see that coprimeness over $\mathbb{Z}_q$ implies the coprimeness over $\mathcal{R}$ [5].

Definition 7 *A polynomial $f(x)$ is said to be basic irreducible over $\mathcal{R}$ if $\psi(f(x)) = \bar{f}(x)$ is irreducible over $\mathbb{Z}_2$.*

Lemma 8 *If $f_1(x), f_2(x), \ldots, f_r(x)$ are r pairwise coprime polynomials over $\mathcal{R}$, then we have*

$$\langle f_1(x)f_2(x)\cdots f_r(x)\rangle = \langle f_1(x)\rangle \cap \langle f_2(x)\rangle \cdots \cap \langle f_r(x)\rangle.$$

Lemma 9 *Let n be an odd positive integer. Then*

1. *$x^n - 1$ has a unique factorization of pairwise coprime polynomials*

$$x^n - 1 = f_1(x)f_2(x)\cdots f_r(x)$$

where each $f_i(x)$ is a basic irreducible polynomial over $\mathcal{R}$ for $i = 1, 2, \ldots, r$.

2. *Let us denote the product of all $f_i(x)$ except $f_j(x)$ by $\hat{f}_j(x)$. Then $f_j(x)$ and $\hat{f}_j(x)$ are coprime for $i = 1, 2, \ldots, r$ and there exist polynomials $u_j(x), v_j(x)$ over $\mathcal{R}$ such that*

$$u_j(x)\hat{f}_j(x) + v_j(x)f_j(x) = 1.$$

3. *Let $e_j = u_j(x)\hat{f}_j(x) + \langle x^n - 1\rangle$ then $1 = e_1 + e_2 + \cdots + e_r$ in $\Re_n$. Moreover $R_j = \Re_n e_j$ is an ideal of $\Re_n$ and hence we have the direct sum decomposition*

$$\Re_n = R_1 \oplus R_2 \oplus \cdots \oplus R_r.$$

4. *For each $1 \leq j \leq r$ the map*

$$\begin{aligned} \mathcal{R}[x]/\langle f_j(x)\rangle \to & R_j = \Re_n e_j \\ k(x) + \langle f_j(x)\rangle \mapsto & (k(x) + \langle x^n - 1\rangle)e_j \end{aligned} \tag{2}$$

is an isomorphism.

Lemma 10 *Let $f(x)$ be a basic irreducible poynomial of degree m over $\mathcal{R}$, where $q = 2^2$ or $q = 2^3$. Then the ideals of $\mathcal{R}[x]/\langle f(x)\rangle$ are the following*

1. *$\langle 2^i + \langle f(x)\rangle\rangle$ for $0 \leq i \leq s$*
2. *$\langle 2^i u + \langle f(x)\rangle\rangle$ for $0 \leq i < s$*
3. *$\langle v(x)2^i + 2^j u + \langle f(x)\rangle\rangle$ where $v(x) = v_0 + v_1 x + \cdots + v_{m-1}x^{m-1}$ is a unit element with $v_i < 2^{min(i-j,s-i)}$ for $i = 0, 1, \ldots, m-1$*

4. $\langle 2^i + \langle f(x)\rangle, 2^j u + \langle f(x)\rangle\rangle$ *for* $0 \le j < i < s$
5. $\langle 2^i + 2^j u + \langle f(x)\rangle, (v-1)2^i + \langle f(x)\rangle\rangle$ *for* $0 \le j < i < s$, $1 < v < 2^{\min(i-j,s-i)}$ *and* $\gcd(2, v) = 1$.

Lemma 11 *The ideals of* $\mathcal{R}[x]/\langle f(x)\rangle$ *given in Lemma 10 are mapped into the ideals of* R_i *as follows*

1. $\langle 2^i \hat{f}_i(x) + \langle x^n - 1\rangle\rangle$ *for* $0 \le i \le s$
2. $\langle 2^i u \hat{f}_i(x) + \langle x^n - 1\rangle\rangle$ *for* $0 \le i < s$ $1 \le v < 2^{\min(i-j,s-i)}$
3. $\langle (v(x)2^i + 2^j u)\hat{f}_i(x) + \langle x^n - 1\rangle\rangle$ *where* $v(x) = v_0 + v_1 x + \cdots + v_{m-1}x^{m-1}$ *is a unit element with* $v_i < 2^{min(i-j,s-i)}$ *for* $i = 0, 1, \ldots, m-1$
4. $\langle 2^i \hat{f}_i(x) + \langle x^n - 1\rangle,\ 2^j u \hat{f}_i(x) + \langle x^n - 1\rangle\rangle$ *for* $0 \le j < i < s$
5. $\langle (2^i + 2^j u)\hat{f}_i(x) + \langle x^n - 1\rangle, (v-1)2^i \hat{f}_i(x) + \langle x^n - 1\rangle\rangle$ *for* $0 \le j < i < s$, $1 < v < 2^{\min(i-j,s-i)}$ *and* $\gcd(2, v) = 1$

under the map given in (2).

Corollary 12 *Let* n *be an odd positive integer and* $f_1 f_2 \cdots f_r$ *be the unique factorization of* $x^n - 1$ *into basic irreducible polynomials and* $q = 2^2$ *or* $q = 2^3$. *Then any ideal of* $\Re_n$ *is a sum of ideals given in Lemma 11*

Corollary 13 *The number of all the ideals in* $\mathcal{R}[x]/\langle f(x)$ *is*

$$K_{s+1} + \binom{s}{2}(2^m - 2)$$

if $q = 2^2$ *or* $q = 2^3$ *where* $f(x)$ *is a basic irreducible polynomial over* $\mathcal{R}$ *of degree* m.

Theorem 14 *Let* $f_1 f_2 \ldots f_r$ *be the unique factorization of* $x^n - 1$ *into basic irreducible polynomials for an odd* n. *Then for* $q = 2^2$ *or* $q = 2^3$ *the number of cyclic codes over* $\mathcal{R}$ *of length* n *is*

$$\begin{aligned}&\prod_{i=1}^{r}\left(K_{s+1} + \binom{s}{2}(2^{m_i} - 2)\right)\\ =&\sum_{i=0}^{r} K_{s+1}^{i}\binom{s}{2}^{r-i}\sum_{1\le j_1<\cdots<j_{r-i}\le r}\ \prod_{k=m_{j_1}}^{m_{j_{r-i}}}(2^k - 2)\end{aligned}$$

where $\deg(f_i(x)) = m_i$.

 ISBN: 978-84-608-6082-2

Now, we give the unique generators for a special family of cyclic codes and their size over $\mathbb{Z}_4 + u\mathbb{Z}_4$.

Theorem 15 *Let n be an odd positive integer and I be an ideal of $(\mathbb{Z}_4 + u\mathbb{Z}_4)[x]/\langle x^n - 1\rangle$ such that $I = I_1 \oplus I_2 \oplus \cdots \oplus I_r$ where each I_i is of the form given in Lemma 11 with $v(x) = 1$ if any one of them is of type 3. Then I is determined by unique monic polynomials f, g, h, s, t, v and z over $\mathbb{Z}_4 + u\mathbb{Z}_4$ such that*

$$I = \langle fhstvz, 2fgstv, ufghtv, 2ufghsvz, (2+u)fghstz\rangle$$

where $fghstvz = x^n - 1$ and

$$|I| = (16)^{deg(g)}(4)^{deg(hsvz^2)}(2)^{deg(t)}.$$

Corollary 16 *Let n be an odd positive integer. Then every ideal of $(\mathbb{Z}_4 + u\mathbb{Z}_4)[x]/\langle x^n - 1\rangle$ given in Theorem 15 is of the form $I = \langle f_0, 2f_1, uf_2, 2uf_3, (2+u)f_4\rangle$, where f_i's are monic divisors of $x^n - 1$ and $f_i | f_0$ over $\mathbb{Z}_4 + u\mathbb{Z}_4$ for $i = 1, 2, 3, 4$ and also $f_3 | f_j$ for $j = 0, 1, 2, 4$.*

Corollary 17 *Let n be an odd positive integer. Then every ideal of $(\mathbb{Z}_4 + u\mathbb{Z}_4)[x]/\langle x^n - 1\rangle$ given in Theorem 15 is of the form $I = \langle f_0 + 2f_1, uf_2 + 2uf_3, (2+u)f_4\rangle$, where f_i's are monic divisors of $x^n - 1$ and $f_i | f_0$ over $\mathbb{Z}_4 + u\mathbb{Z}_4$ for $i = 1, 2, 3, 4$ and also $f_3 | f_j$ for $j = 0, 1, 2, 4$.*

4 Conclusion

In this work, we determined the ideal structure of the ring $\mathbb{Z}_q + u\mathbb{Z}_q$ where q is a power of 2 and $u^2 = 0$. We also attempted to determine the ideal structure of $(\mathbb{Z}_q + u\mathbb{Z}_q)/\langle f(x)\rangle$ where $f(x)$ is a basic irreducible polynomial over $\mathbb{Z}_q + u\mathbb{Z}_q$ for the purpose of determining the ideal structure of $(\mathbb{Z}_q + u\mathbb{Z}_q)[x]/\langle x^n - 1\rangle$ and hence obtaining cyclic codes of length n which is an odd number. We achieved this goal completely for $q = 2^2$ and $q = 2^3$. However, for q^s where $s > 3$, it is still an open problem. We gave the definite generators for a family of some cyclic codes over $\mathbb{Z}_4 + u\mathbb{Z}_4$ by divisors of $x^n - 1$ and the size of these cyclic codes.

References

[1] A. Bonnecaze, U. Prampalli, Cyclic codes and self-dual codes over $\mathbb{F}_2 + u\mathbb{F}_2$, *Information Theory, IEEE Transactions on*, **45** (1999), 1250–1255.

[2] B. Yildiz and N. Aydin, On cyclic codes over $\mathbb{Z}_4 + u\mathbb{Z}_4$ and their $\mathbb{Z}_4-$images, *International Journal of Information and Coding Theory*, **2** (2014), 226–237.

[3] D. S. Dummit and R. M. Foote, *Abstract algebra*, Englewood Cliffs: Prentice Hall,1991.

[4] E. Prange, *Cyclic Error-Correcting codes in two symbols*, Air Force Cambridge Research Center, 1957.

[5] J. Gao, F.W. Fu, L. Xiao and R. Bandi, Some results on cyclic codes over $\mathbb{Z}_q + u\mathbb{Z}_q$, *Discrete Mathematics, Algorithms and Applications*, **7** (2015).

[6] J. W. Layman, "The On-Line Encyclopedia of Integer Sequences, Sequence A053599".

[7] M. Al-Ashker and H. Mohammed, Cyclic codes over $\mathbb{Z}_2 + u\mathbb{Z}_2 + u^2\mathbb{Z}_2 + \ + u^{k-1}\mathbb{Z}_2$, *Turkish J. Math*, **35** (2011), 737–749.

[8] Q. Jian-Fa, Z. Li-na and Z. Shi-Xin, Cyclic Codes over $\mathbb{F}_p + u\mathbb{F}_p + ... + u^{k-1}\mathbb{F}_p$, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **2** (2014), 226–237.

[9] R. Hammons, V. P. Kumar, A. R. Calderbank, N. Sloane and P. Sole, The Z 4-linearity of Kerdock, Preparata, Goethals, and related codes, *Information Theory, IEEE Transactions on*, **40** (1994), 301–319.

[10] T. Abualrub and I. Siap, Cyclic codes over the rings $\mathbb{Z}_2 + u\mathbb{Z}_2$ and $\mathbb{Z}_2 + u\mathbb{Z}_2 + u^2\mathbb{Z}_2$, *Designs, Codes and Cryptography*, **43** (2007), 273–287.

[11] Z. Wan, *Quaternery Codes*, 2^{nd} edition, World Scientific, Singapore, 1997.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

A DRBEM approach for the Stokes eigenvalue problem

M. Tezer-Sezgin[1] **and Önder Türk**[2]

[1] *Department of Mathematics , Middle East Technical University and*

[2] *Department of Mathematics , Gebze Technical University*

emails: munt@metu.edu.tr, onder.turk@yandex.com

Abstract

This study considers an approximation to the Stokes eigenvalue problem based on the dual reciprocity boundary element method (DRBEM). The momentum equations in primitive variables, namely, displacement and pressure, are cross-differentiated and subtracted to eliminate the pressure gradient terms. Introducing stream function and vorticity, results in a system composed of the stream function equation and the vorticity eigenvalue equation, also automatically satisfies the continuity equation. The vorticity eigenvalue problem now has the same eigenvalues of the Stokes operator. The DRBEM approach consists in weighting the vorticity eigenvalue equation with the fundamental solution of the Laplace equation, where the terms except Laplacian are treated as inhomogeneity. The domain integrals arising from the source terms are transformed into a series of boundary integrals requiring that the right hand side of the equations be approximated by a radial basis function interpolation. In the present work, the collocation nodes are distributed on the boundary and in the interior of the problem domain under consideration. The matrices resulting from the application of DRBEM to the vorticity eigenvalue equation are partitioned into submatrices to eliminate the fluxes of the vorticity on the boundary. The stream function equation is approximated by the coordinate matrix constructed from the radial basis function at the boundary and interior nodes. The vorticity boundary and interior values are inserted from DRBEM discretized system of the vorticity eigenvalue equation, and the homogeneous stream function boundary values are dropped. This way, the system of two equations is reduced into a single eigenvalue equation in terms of the stream function partition corresponding to the interior nodes. Moreover, the resulting eigenvalue equation is converted to a generalized eigenvalue problem in order to avoid any matrix inversion for the sake of numerical efficiency. The boundary is discretized by using constant elements, leading to a small algebraic system solved at a small expense compared to domain discretization methods. The method is first tested on a Laplace eigenvalue problem with known analytical solution, and the results are validated by a comparison of several approximations with the corresponding exact eigenpairs. Furthermore, the numerical results obtained, verify the numerical convergence to the reference eigenvalues of the Stokes operator and to the corresponding eigenfunctions.

Key words: DRBEM, Stokes eigenvalue problem

1 Introduction

The Stokes eigenvalue problem is a subject of an active research due to its significance on both fundamental and practical grounds. The eigenmodes of the Stokes operator constitute a natural basis to analyze the homogeneous component of any flow, in particular, for describing the fluctuating part of a turbulent flow. Moreover, the Stokes eigenvalue problem is used as benchmark for analyzing the convergence and accuracy for numerical algorithms designed in fluid dynamics [1, 2].

There are numerous recent works approximating the Stokes eigenproblem, most of which are based on finite elements (see e.g. [3, 4, 5, 6, 7]). On the other hand, a mesh free method based on radial basis functions is developed to approximate the eigenvalues of the Stokes operator given in primitive variables on a square domain in [8]. Concerning the source problem, an iterative dual reciprocity boundary element method (DRBEM) based on the compactly-supported, positive definite radial basis function for the solution of Stokes flow problems is considered in [9].

This study considers a first DRBEM approximation to the Stokes eigenvalue problem to the best of authors' knowledge. The governing equations are transformed into stream function and vorticity form, resulting in the vorticity eigenvalue equation which has the same eigenvalues as the Stokes operator, and an additional stream function equation. The DRBEM approach consists in weighting the vorticity eigenvalue equation with the fundamental solution of the Laplace equation, where the terms except Laplacian are treated as inhomogeneity. The domain integrals arising from the source terms are transformed into a series of boundary integrals requiring that the right hand side of the equations be approximated by a radial basis function interpolation. Apart from resulting in a considerable reduction in terms of the computational work compared to domain discretization methods, the method also has the advantage of calculating all the space derivatives using DRBEM coordinate matrix. The matrices obtained in the vorticity eigenvalue equation are partitioned into submatrices to eliminate the fluxes of the vorticity on the boundary. The stream function equation is also approximated by the DRBEM coordinate matrix constructed from the radial basis function at the boundary and interior nodes. The vorticity boundary and interior values are inserted from the solution of the vorticity eigenvalue equation, and then the homogeneous stream function boundary values are dropped. Thus, the system of two equations is reduced into a single eigenvalue equation in terms of the stream function partition corresponding to the interior nodes. Moreover, the resulting eigenvalue equation is converted to a generalized eigenvalue problem in order to avoid any matrix inversion for the sake of numerical efficiency. The boundary is discretized by using constant elements, leading to a small algebraic system solved at a small expense.

2 Governing equations

Let Ω be the bounded and polyhedral computational domain in $\mathbb{R}^2$, and $\partial\Omega$ be its boundary. The Stokes eigenvalue problem is considered as follows: find $[\mathbf{u}, p, \lambda]$, where $\mathbf{u} \neq 0$ is the displacement

or velocity field, p is the pressure, and $\lambda \in \mathbb{R}$, such that [3]

$$\begin{cases} -\mu\Delta\mathbf{u} + \nabla p = \lambda\mathbf{u} & \text{in } \Omega, \\ \nabla\cdot\mathbf{u} = 0 & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega. \end{cases} \tag{1}$$

The first equation in (1) can be written component-wise in two-dimensional form as

$$\begin{aligned} -\mu\Delta u + \frac{\partial p}{\partial x} &= \lambda u, \\ -\mu\Delta v + \frac{\partial p}{\partial y} &= \lambda v, \end{aligned} \tag{2}$$

where u and v are the velocity components of the displacement vector $\mathbf{u}$. The constraint of continuity takes the form

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0,$$

and the boundary conditions for the velocity components are given as $u = 0$ and $v = 0$.

The stream function ψ satisfying the continuity equation directly, is defined in 2D as

$$\frac{\partial \psi}{\partial y} = u, \quad \frac{\partial \psi}{\partial x} = -v. \tag{3}$$

Then, the only nonzero component of the vorticity field is

$$w = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = -\left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2}\right) = -\Delta\psi. \tag{4}$$

The direct solution to System (2) has well known difficulties due to the pressure term. Therefore, the well established stream function-vorticity formulation which is used extensively in the literature for solving incompressible fluid flow problems is considered. The pressure term is eliminated by subtracting the derivative of the first equation in (2) with respect to y from the derivative of the second equation with respect to x. Thus, a system composed of an equation standing for the stream function, and the vorticity eigenvalue equation is obtained as follows

$$\begin{aligned} -\mu\Delta w &= \lambda w, \\ \Delta\psi &= -w. \end{aligned} \tag{5}$$

The solution of vorticity eigenvalue equation requires the stream function equation to be solved in the above system. The vorticity eigenvalue problem now has the same eigenvalues of the Stokes operator. $\psi = 0$ is the boundary condition for the stream function, however, the boundary conditions of the vorticity are not known. A DRBEM approach for solving (5) also proposing a remedy to this deficiency is presented in the next section.

3 DRBEM formulation

The aim of the DRBEM is to transform the governing equations of the problem into boundary integral equations. It consists in weighting equations with the fundamental solution of the Laplace equation, where the terms except Laplacian are treated as inhomogeneity. The domain integrals arising from the source terms are transformed into boundary integrals approximating the inhomogeneity by radial basis functions f_j which are related to particular solutions $\hat{u}_j$ with $\nabla^2 \hat{u}_j = f_j$ [10].

The equations in (5) are weighted with the two-dimensional fundamental solution of Laplace equation, $u^* = 1/2\pi \ln(1/r)$, and application of the Green's second identity results in

$$\begin{aligned} c_i \psi_i + \int_\Gamma (q^* \psi - u^* \frac{\partial \psi}{\partial n}) d\Gamma &= -\int_\Omega (-w) u^* d\Omega, \\ c_i w_i + \int_\Gamma (q^* w - u^* \frac{\partial w}{\partial n}) d\Gamma &= -\int_\Omega (\lambda w) u^* d\Omega, \end{aligned} \tag{6}$$

where $q^* = \partial u^*/\partial n$, Γ is the boundary of the domain Ω, and the subscript i denotes the source point. The constant c_i is given by $c_i = \theta_i/2\pi$ with the internal angle θ_i at the source point. The approximations for the integrands in the domain integrals are given by $\sum_{j=1}^{N+L} \alpha_j f_j(x,y)$ and $\sum_{j=1}^{N+L} \beta_j f_j(x,y)$, the coefficients α_j and β_j being undetermined constants. The numbers of the boundary and the internal nodes are denoted by N and L, respectively. The substitution of $\nabla^2 \hat{u}_j = f_j$ enables to use BEM approach also to the domain integral which results in boundary integrals only. The use of constant elements for the discretization of the boundary gives then the corresponding matrix-vector equations

$$\begin{aligned} H\psi - G\frac{\partial \psi}{\partial n} &= (H\hat{U} - G\hat{Q})F^{-1}(-w) = -Sw, \\ Hw - G\frac{\partial w}{\partial n} &= (H\hat{U} - G\hat{Q})F^{-1}(\lambda w) = \lambda Sw, \end{aligned} \tag{7}$$

where the matrices $\hat{U}$ and $\hat{Q}$ are constructed by taking each of the vectors $\hat{u}_j$ and $\hat{q}_j$ as columns, respectively.

The components of the matrices H and G are

$$H_{ij} = c_i \delta_{ij} + \frac{1}{2\pi} \int_{\Gamma_j} \frac{\partial}{\partial n} \left(\ln(\frac{1}{r}) \right) d\Gamma_j, \quad H_{ii} = -\sum_{j=1, j\neq i}^{N} H_{ij}, \quad G_{ij} = \frac{1}{2\pi} \int_{\Gamma_j} \ln(\frac{1}{r}) d\Gamma_j, \quad G_{ii} = \frac{A}{2\pi}(\ln(\frac{2}{A}) + 1),$$

where r is the distance from node i to element j, A is the length of the element and δ_{ij} is the Kronecker delta function. The coordinate matrix F of size $(N+L)$ contains the radial basis functions f_j as columns. In this study, the compactly supported positive definite radial basis functions are made use of in the form of a univariable polynomial of minimal degree for a given dimension and smoothness,

$$\phi(r) = \begin{cases} \left(1 - \frac{r}{a}\right)^4 \left(1 + \frac{4r}{a}\right), & \text{for } 0 \le a \le r, \\ 0, & \text{for } r > a, \end{cases}$$

 ISBN: 978-84-608-6082-2

where a is an influence radius truncating the function to zero [9, 11].

The matrices resulting from the application of DRBEM to the stream function equation are partitioned into submatrices following the procedure given in [10, 12], to eliminate the fluxes of the vorticity on the boundary, as

$$\begin{bmatrix} H_{\mathrm{bb}} & H_{\mathrm{bi}} \\ H_{\mathrm{ib}} & H_{\mathrm{ii}} \end{bmatrix} \begin{bmatrix} \psi_{\mathrm{b}} \\ \psi_{\mathrm{i}} \end{bmatrix} - \begin{bmatrix} G_{\mathrm{bb}} & G_{\mathrm{bi}} \\ G_{\mathrm{ib}} & G_{\mathrm{ii}} \end{bmatrix} \begin{bmatrix} \frac{\partial \psi}{\partial n}|_{\mathrm{b}} \\ 0 \end{bmatrix} = \begin{bmatrix} S_{\mathrm{bb}} & S_{\mathrm{bi}} \\ S_{\mathrm{ib}} & S_{\mathrm{ii}} \end{bmatrix} \begin{bmatrix} w_{\mathrm{b}} \\ w_{\mathrm{i}} \end{bmatrix}. \tag{8}$$

Similarly, the vorticty eigenvalue equation is written as

$$\begin{bmatrix} H_{\mathrm{bb}} & H_{\mathrm{bi}} \\ H_{\mathrm{ib}} & H_{\mathrm{ii}} \end{bmatrix} \begin{bmatrix} w_{\mathrm{b}} \\ w_{\mathrm{i}} \end{bmatrix} - \begin{bmatrix} G_{\mathrm{bb}} & G_{\mathrm{bi}} \\ G_{\mathrm{ib}} & G_{\mathrm{ii}} \end{bmatrix} \begin{bmatrix} \frac{\partial w}{\partial n}|_{\mathrm{b}} \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} S_{\mathrm{bb}} & S_{\mathrm{bi}} \\ S_{\mathrm{ib}} & S_{\mathrm{ii}} \end{bmatrix} \begin{bmatrix} w_{\mathrm{b}} \\ w_{\mathrm{i}} \end{bmatrix}. \tag{9}$$

$\frac{\partial \psi}{\partial n}|_{\mathrm{b}}$ in (8) and $\frac{\partial w}{\partial n}|_{\mathrm{b}}$ in (9) can be eliminated to obtain

$$\psi_{\mathrm{i}} = A w_{\mathrm{b}} + B w_{\mathrm{i}} \quad \text{and} \quad C w_{\mathrm{b}} + w_{\mathrm{i}} = \lambda A w_{\mathrm{b}} + \lambda B w_{\mathrm{i}} \tag{10}$$

where

$$A = S_{\mathrm{ib}} - G_{\mathrm{ib}} G_{\mathrm{bb}}^{-1} S_{\mathrm{bb}}, \quad B = S_{\mathrm{ii}} - G_{\mathrm{ib}} G_{\mathrm{bb}}^{-1} S_{\mathrm{bi}} \quad \text{and} \quad C = H_{\mathrm{ib}} - G_{\mathrm{ib}} G_{\mathrm{bb}}^{-1} H_{\mathrm{bb}}.$$

The two systems in (10) are combined as

$$(C - B^{-1}A) w_{\mathrm{b}} + B^{-1} \psi_{\mathrm{i}} = \lambda \psi_{\mathrm{i}}, \tag{11}$$

involving the original eigenvalue λ, and two unknown vectors ψ_{i} and w_{b}. The strategy is to transform this equation into one eigenvalue problem in terms of stream function vector ψ_{i}. To achieve this, the stream function equation in (5) is discretized by the coordinate matrix F as

$$w = D\psi, \quad D = -\left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2}\right) F^{-1}.$$

Making the same partition for ψ_{i}, ψ_{b} and w_{b} one can obtain the system

$$w_{\mathrm{b}} = D_{\mathrm{bi}} \psi_{\mathrm{i}}.$$

Finally, substituting this system in (11), the eigenvalue equation in terms of ψ_{i} only is obtained as

$$\left[B^{-1} + (C - B^{-1}A) D_{\mathrm{bi}}\right] \psi_{\mathrm{i}} = \lambda \psi_{\mathrm{i}}. \tag{12}$$

When the above equation is left multiplied by B, it is converted to a generalized eigenvalue problem,

$$\left[I + (BC - A) D_{\mathrm{bi}}\right] \psi_{\mathrm{i}} = \lambda B \psi_{\mathrm{i}}, \tag{13}$$

where I is the identity matrix, and thus any matrix inversion is avoided for the sake of numerical efficiency. Obviously, the eigenvalues of (13) are the same as the ones of (1) and (2), where the corresponding eiqenvector is the approximation to the stream function interior partition.

 ISBN: 978-84-608-6082-2

4 Numerical results

In this section, numerical comparisons are presented to show the efficiency of the approach proposed. The technique is first tested on a Laplace eigenvalue problem in a square domain with known analytical solution, and the results are validated by a comparison of several approximations with the corresponding exact eigenpairs. Next, the Stokes eigenvalue problem in considered. The convergence behaviors are examined for the reference eigenvalue approximations in terms of the relative error given as the ratio of the difference between the approximate value and the reference value, to the latter. The computations are carried out for several N, the number of constant boundary elements, and the influence radius is taken as $a = 0.5$. The well known Chebyshev-Gauss-Lobatto points are used as collocation points in the interior of the problem domain due to their desired property of uneven distribution clustering near the boundaries. Since $N = L$, the total number of nodes is $2N$ in each case.

4.1 Laplace eigenvalue problem

As a first test, the two-dimensional Laplace eigenvalue problem is considered: find eigenvalues λ and nontrivial eigenfunctions u such that

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{14}$$

on the square $\Omega = [0,\pi] \times [0,\pi]$. The eigenvalues of (14) are given as $\lambda_{m,n} = m^2 + n^2$ where m and n are positive integers, and the corresponding eigenfunctions are $u_{m,n} = \sin(mx)\sin(ny)$. The first eigenvalue $\lambda_{1,1} = 2$ is taken as the reference solution, and the corresponding eigenvector $u_{1,1} = \sin(x)\sin(y)$. is approximated. The computed eigenvalues with the relative errors are presented in Table 1. The results show that as N increases, the approximate value converges to the exact value. The contours of the corresponding eigenvector approximation u_h together with the exact eigenfunction $u_{1,1}$ are illustrated in Figure 1 where the perfect agreement between them is well observed.

 ISBN: 978-84-608-6082-2

Table 1: Computed eigenvalues of the Laplace eigenvalue problem.

N	λ_h	$(\lambda_h - \lambda_{1,1})/\lambda_{1,1}$
16	2.640578420197327	0.320289210098663
25	2.046959119321968	0.023479559660984
36	2.001893432202034	0.000946716101017
49	2.000107494692682	0.000053747346341
64	2.000012259988895	0.000006129994448
81	2.000001479988846	0.000000739994423

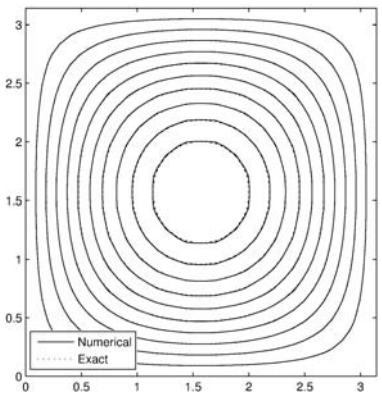


Figure 1: Computed and exact eigenfunctions of the Laplace eigenvalue problem, $N = 25$.

4.2 Stokes eigenvalue problem

The numerical solution to the Stokes eigenvalue problem by means of (13) in $\Omega = [0,1] \times [0,1]$ is presented in this part. The results are presented for the approximations to the first eigenvalue with the reference value $\lambda_1 = 52.3446911$ [8]. The approximations to the first eigenvalue as well as the error values, are listed in Table 2. It can be seen from this table that as the number of elements (N) increase, the computed eigenvalue converges to the reference value monotonically from above.

In order to compare the results qualitatively, the plot of the displacement vector, when $N = 25$ is given in Figure 2. The figure puts forward the good agreement between the behaviors of the velocity streamlines (flow vector) and the previously published results [3, 5].

Table 2: Computed eigenvalues of the Stokes eigenvalue problem.

N	λ_h	$(\lambda_h - \lambda_1)/\lambda_1$
16	69.321824666669457	0.324333436875883
25	57.945342749714712	0.106995600356398
36	53.917133782677567	0.030040155928584
49	53.045969289374028	0.013397312595356
64	52.563251505171714	0.004175407296877
81	52.344756214734844	0.000001243960629

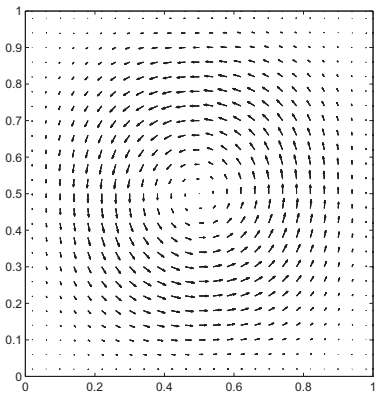


Figure 2: Flow vector of the Stokes eigenvalue problem, $N = 25$.

5 Conclusion

The DRBEM formulation based on the partition technique is presented to solve the Stokes eigenvalue problem. The stream function and vorticity formulation is considered, resulting in two equations, namely, a stream function equation and a vorticity eigenvalue equation. The matrices resulted from the application of DRBEM are partitioned into submatrices, and the fluxes on the boundary are eliminated. The system of two equations is reduced into a single eigenvalue equation in terms of the stream function partition that corresponds to the interior nodes. The boundary is discretized by using constant elements which results in a small algebraic system solved at a small expense. The coordinate matrix is generated by using the compactly supported positive definite radial basis functions in the form of a univariable polynomial of minimal degree for a given dimension and smoothness.

Moreover, the Chebyshev-Gauss-Lobatto points are used as collocation points in the interior of the problem domain with the property of uneven distribution clustering near the boundaries. The proposed technique is first validated by means of a test problem, the Laplace eigenproblem on a square domain. The approximate values are shown to be converging to the reference eigenvalue, and in addition, it is set forth that the contours of the approximate eigenvector possess the same behavior as those of the exact solution. The same approach is applied to approximate the eigenvalues of the Stokes operator. The numerical results obtained, verify the numerical convergence to the reference eigenvalue. The displacement vector is also shown to be in good agreement with the ones provided in the literature.

References

[1] G. LABROSSE, E. LERICHE AND P. LALLEMAND, *Stokes eigenmodes in cubic domain: their symmetry properties*, Theoretical and Computational Fluid Dynamics **28** (2014) 335–256.

[2] KAI SCHNEIDER AND MARIE FARGE, *Final states of decaying 2D turbulence in bounded domains: Influence of the geometry*, Physica D: Nonlinear Phenomena **237** (2008) 2228–2233.

[3] PENGZHAN HUANG, YINNIAN HE, AND XINLONG FENG, *Numerical investigations on several stabilized finite element methods for the Stokes eigenvalue problem*, Mathematical Problems in Engineering **Article ID 745908 2011** (2011) 1–14.

[4] HUIPO LIU, WEI GONG, SHUANGHU WANG, AND NINGNING YAN, *Superconvergence and a posteriori error estimates for the stokes eigenvalue problems*, BIT Numerical Mathematics **53** (2013) 665–687.

[5] MARIA G. ARMENTANO AND VERONICA MORENO, *A posteriori error estimates of stabilized low-order mixed finite elements for the Stokes eigenvalue problem*, Journal of Computational and Applied Mathematics **269** (2014) 132–149.

[6] HEHU XIE AND XIAOBO YIN, *Acceleration of stabilized finite element discretizations for the Stokes eigenvalue problem*, Advances in Computational Mathematics **41** (2015) 799–812.

[7] PENGZHAN HUANG, *Lower and upper bounds of Stokes eigenvalue problem based on stabilized finite element methods*, Calcolo **52** (2015) 109–121.

[8] A. GOLBABAI AND H. RABIEI, *A meshfree method based on radial basis functions for the eigenvalues of transient stokes equations*, Engineering Analysis with Boundary Elements, **36** (2012) 1555–1559.

[9] D. L. YOUNG, C. C. TSAI, T. I. ELDHO, AND A. H. -D. CHENG, *Solution of stokes flow using an iterative DRBEM based on compactly-supported, positive-definite radial basis function*, Computers and Mathematics with Applications **43** (2002) 607–619.

[10] P. W. PARTRIDGE, C. A. BREBBIA, AND L. C. WROBEL, *The Dual Reciprocity Boundary Element Method*, Southampton, Boston, 1992.

[11] A. H. -D. CHENG, D. -L. YOUNG, AND C. -C. TSAI, *Solution of poissons equation by iterative DRBEM using compactly supported, positive definite radial basis function*, Engineering Analysis with Boundary Elements **24** (2000) 549–557.

[12] D. NARDINI AND C. A. BREBBIA, *A new approach to free vibration analysis using boundary elements*, Applied Mathematical Modelling **7** (1983) 157–162.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Bias-induced effects in single molecule transport

Jos Thijssen[1], Jose Celis Gil[1] and Josko de Boer[1]

[1] *Kavli Institute of Nanoscience, Delft University of Technology*

emails: J.M.Thijssen@tudelft.nl, J.A.CelisGil@tudelft.nl,
joskodeboer@gmail.com

Abstract

In a few recent papers [1, 2], bias-induced phenomena were studied, in which intramolecular transport was explained by us in terms of resonances. Although nonequilibrium Green's functions applied to theoretical models and in quantum chemistry calculations showed beautiful agreement with experiment, there were also important discrepancies, notably in the current values and in the location of the Fermi energy. We discuss recent progress on the theory, which shows that inclusion of Coulomb interactions can account for the discrepancies. Not only does the theory yield current values in agreement with the experiment, also the apparent alignment of the molecular HOMO levels with the Fermi energy, observed in [1], are reproduced when using realistic Coulomb parameters.

We also discuss methods for locating the molecular frontier orbitals with respect to the Fermi energy, relying on spin-dependent transport.

Key words: Single molecule charge transport, nonequilibrium Green's function techniques

Acknowledgements

This research was performed with financial support from The Netherlands Organization for Scientific Research (NWO/ OCW), FOM.

References

[1] Mickael L Perrin, Riccardo Frisenda, Max Koole, Johannes S Seldenthuis, Jose A Celis Gil, Hennie Valkenier, Jan C Hummelen, Nicolas Renaud, Ferdinand C Grozema, Joseph M Thijssen, et al. Large negative differential conductance in single-molecule break junctions. *Nature nanotechnology*, 9(10):830–834, 2014.

[2] Mickael L Perrin, Elena Galán, Rienk Eelkema, Joseph M Thijssen, Ferdinand Grozema, and Herre SJ van der Zant. A gate-tunable single-molecule diode. *Nanoscale*, 8(16):8919–8923, 2016.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Factorization and inversion of finite and infinite bordered tridiagonal matrices

Venancio Tomeo[1]

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

emails: tomeo@ucm.es

Abstract

Motivated by current applications of the bordered tridiagonal matrices of large order, the infinite invertible bordered tridiagonal matrices are considered. These matrices are a generalization of the tridiagonal and the arrowhead matrices. A method based on a simple but suitable factorization is proposed for obtaining, in the finite-dimensional case, a decomposition of the inverse. This procedure can be applicable to infinite bordered tridiagonal matrices. Some illustrative examples are given.

Key words: Bordered tridiagonal matrices, inverse matrix, infinite matrix, tridiagonal matrix.

1 Introduction

Tridiagonal matrices, arrowhead matrices and bordered tridiagonal matrices frequently appear in scientific and engineering investigations and computational physics; e.g. matrix algebra, numerical solutions of elliptic differential equations, interpolation by spline functions, electronic circuit simulation, fluid flow problems, heat conduction, boundary value problems, telecommunications system analysis, parallel computing, see [5, 8, 11, 12, 14] and the references given there.

Tridiagonal and related matrices has been extensely studied, see for example [1, 2, 6, 7, 9, 10]. The bordered tridiagonal matrices generalize the tridiagonal and the arrowhead matrices. It motivates our study regarding the factorization and the inversion of general bordered tridiagonal matrices. Previous works can be seen in [10] and some references therein.

For inverting an $n \times n$ bordered tridiagonal matrix, instead of the border lines at the first row and column, some authors consider the border lines at the last row and column, because we can obtain one form from the other with a permutation matrix $P = (p_{ij})$, where $p_{ij} = 1$ if $i + j = n$, and zero otherwise. Since our interest is the study of the infinite case, we consider the border lines at the first row and column.

The material is organized as follows. In Section 2, we show basic results about the UTL factorization for the finite bordered tridiagonal matrices. Section 3 is dedicated to particular cases of finite and infinite tridiagonal matrices. In Section 4 we study the factorization and inversion of infinite bordered tridiagonal matrices. Throughout the text the results are illustrated with appropriate examples.

2 Finite bordered tridiagonal matrices

An $n \times n$ real or complex matrix A_n is called a *bordered tridiagonal matrix* if A_n is of the form

$$A_n = \left(\begin{array}{c|ccccc} b_0 & u_1 & u_2 & \cdots & u_{n-2} & u_{n-1} \\ \hline l_1 & b_1 & c_2 & & & \\ l_2 & a_2 & b_2 & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ l_{n-2} & & & \ddots & b_{n-2} & c_{n-1} \\ l_{n-1} & & & & a_{n-1} & b_{n-1} \end{array}\right). \tag{1}$$

The idea is to factorize the bordered tridiagonal matrix A_n in the form UTL, where U is an upper triangular matrix, L is a lower triangular matrix and T is a tridiagonal matrix. That is, $A_n = U_n T_n L_n =$

$$\left(\begin{array}{cccccc} 1 & u_1' & u_2' & \cdots & u_{n-2}' & u_{n-1}' \\ & 1 & c & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{array}\right) \left(\begin{array}{cccccc} b_0' & c_1' & & & & \\ a_1' & b_1 & c_2 & & & \\ & a_2 & b_2 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & b_{n-2} & c_{n-1} \\ & & & & a_{n-1} & b_{n-1} \end{array}\right) \left(\begin{array}{cccccc} 1 & & & & & \\ l_1' & 1 & & & & \\ l_2' & & 1 & & & \\ \vdots & & & \ddots & & \\ l_{n-2}' & & & & 1 & \\ l_{n-1}' & & & & & 1 \end{array}\right).$$

Matrix A_n has $5n-6$ nonzero entries and the product of the three matrices has $5n-4$ nonzero entries. Then, we can choose two entries en matrix T_n, for example $a_1' = c_1' = 0$, and determinate the other entries.

Matrix $\widehat{T}_{n-1}$ denote the tridiagonal matrix obtained from A_n, or from T_n, when first row and first column are deleted. Now we present the Theorem for the finite case.

Theorem 1. *Every nonsingular bordered tridiagonal matrix A_n as given in (1), with* $\det \widehat{T}_{n-1} \neq 0$*, has an unique UTL factorization of the form* $A_n = U_n T_n L_n =$

$$\left(\begin{array}{cccccc} 1 & u'_1 & u'_2 & \cdots & u'_{n-2} & u'_{n-1} \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{array}\right) \left(\begin{array}{c|ccccc} b'_0 & & & & & \\ \hline & b_1 & c_2 & & & \\ & a_2 & b_2 & c_3 & & \\ & & a_3 & \ddots & \ddots & \\ & & & \ddots & b_{n-2} & c_{n-1} \\ & & & & a_{n-1} & b_{n-1} \end{array}\right) \left(\begin{array}{cccccc} 1 & & & & & \\ l'_1 & 1 & & & & \\ l'_2 & & 1 & & & \\ \vdots & & & \ddots & & \\ l'_{n-2} & & & & 1 & \\ l'_{n-1} & & & & & 1 \end{array}\right),$$

where the entries can be evaluated, from the inverse matrix of $\widehat{T}_{n-1}$*, by*

$$\begin{pmatrix} u'_1 \\ u'_2 \\ \vdots \\ u'_{n-1} \end{pmatrix} = \left(\widehat{T}_{n-1}^{-1}\right)^t \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{pmatrix}, \qquad \begin{pmatrix} l'_1 \\ l'_2 \\ \vdots \\ l'_{n-1} \end{pmatrix} = \widehat{T}_{n-1}^{-1} \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_{n-1} \end{pmatrix} \text{ and } b'_0 = \frac{\det A_n}{\det \widehat{T}_{n-1}}.$$

Proof. Taking the product $U_n T_n L_n$ and identifying it with matrix A_n, we obtain the systems

$$\begin{cases} u_1 = u'_1 b_1 + u'_2 a_2 \\ u_2 = u'_1 c_2 + u'_2 b_2 + u'_3 a_3 \\ u_3 = \quad u'_2 c_3 + u'_3 b_3 + u'_4 a_4 \\ \quad \vdots \\ u_i = \qquad u'_{i-1} c_i + u'_i b_i + u'_{i+1} a_{i+1} \end{cases} \qquad \begin{cases} l_1 = l'_1 b_1 + l'_2 c_2 \\ l_2 = l'_1 a_2 + l'_2 b_2 + l'_3 c_3 \\ l_3 = \quad l'_2 a_3 + l'_3 b_3 + u'_4 c_4 \\ \quad \vdots \\ l_i = \qquad l'_{i-1} a_i + l'_i b_i + l'_{i+1} c_{i+1}, \end{cases}$$

for $i = 1, 2, ...,$ with $a_1 = c_1 = a_n = c_n = 0$. Then, we have

$$\begin{aligned} u'_{i+1} &= \frac{u_i - u'_{i-1} c_i - u'_i b_i}{a_{i+1}} \\ l'_{i+1} &= \frac{l_i - l'_{i-1} a_i - l'_i b_i}{c_{i+1}}, \qquad i = 1, 2, ..., n-1, \\ b'_0 &= b_0 - \sum_{i=1}^{n-1} l'_i (u'_i c_i + u'_{i+1} b_i + u'_{i+1} a_{i+1}) = \frac{\det A_n}{\det \widehat{T}_{n-1}}. \end{aligned}$$

The systems can be solved by

$$(u'_1, u'_2, \ldots, u'_n) = (u_1, u_2, \ldots, u_n) \cdot \widehat{T}_{n-1}^{-1}.$$

That is

$$\begin{pmatrix} u'_1 \\ u'_2 \\ \vdots \\ u'_n \end{pmatrix} = \left(\widehat{T}_{n-1}^{-1}\right)^t \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} l'_1 \\ l'_2 \\ \vdots \\ l'_n \end{pmatrix} = \widehat{T}_{n-1}^{-1} \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{pmatrix}.$$

 ISBN: 978-84-608-6082-2

The proof for uniqueness is simple: if U', T' and L' are matrices as in the establishment, taking the product $U'T'L' =$

$$\begin{pmatrix} 1 & u_1'' & u_2'' & \cdots & u_{n-2}'' & u_{n-1}'' \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix} \left(\begin{array}{c|cccc} b_0'' & & & & \\ \hline & b_1 & c_2 & & \\ & a_2 & b_2 & c_3 & \\ & & a_3 & \ddots & \ddots \\ & & & \ddots & b_{n-2} & c_{n-1} \\ & & & & a_{n-1} & b_{n-1} \end{array}\right) \begin{pmatrix} 1 & & & & & \\ l_1'' & 1 & & & & \\ l_2'' & & 1 & & & \\ \vdots & & & \ddots & & \\ l_{n-2}'' & & & & 1 & \\ l_{n-1}'' & & & & & 1 \end{pmatrix},$$

as matrix $\widehat{T}_{n-1}$ is the same, then we have $u_i'' = u_i'$, $l_i'' = l_i'$, $k_i' = k_i$, and $b_0'' = b_0$. Thus the proposed factorization is unique. □

When some a_i or c_i are zero, because matrix $\widehat{T}_{n-1}$ is nonsingular, the entries u_i', l_i' and b_0' can be evaluated. In particular, if $a_i = c_i = 0$, for $i = 1, 2, ..., n-1$, the matrix $\widehat{T}_{n-1}$ is diagonal. This case has been studied as *arrowhead matrices* in [3]. Now, we present two trivial corollaries.

Corollary 1. *The inverse of a nonsingular bordered tridiagonal matrix A_n satisfying Theorem 1 has a factorization of the form $A_n^{-1} = L_n^{-1}T_n^{-1}U_n^{-1} =$*

$$= \begin{pmatrix} 1 & & & & & \\ -l_1' & 1 & & & & \\ -l_2' & & 1 & & & \\ \vdots & & & \ddots & & \\ -l_{n-2}' & & & & 1 & \\ -l_{n-1}' & & & & & 1 \end{pmatrix} \left(\begin{array}{c|ccc} \frac{1}{b_0'} & & & \\ \hline & & & \\ & & \widehat{T}_{n-1}^{-1} & \\ & & & \end{array}\right) \begin{pmatrix} 1 & -u_1' & -u_2' & \cdots & -u_{n-2}' & -u_{n-1}' \\ & 1 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix}.$$

Corollary 2. *The inverse A_n^{-1} of a nonsingular bordered tridiagonal matrix A_n satisfying Theorem 1 can be decomposed as the sum of the inverse of a tridiagonal plus a rank-one matrix in the form*

$$A_n^{-1} = \left(\begin{array}{c|ccc} 0 & & & \\ \hline & & & \\ & & \widehat{T}_{n-1}^{-1} & \\ & & & \end{array}\right) + \frac{1}{b_0'} \begin{pmatrix} 1 \\ -l_1' \\ -l_2' \\ \vdots \\ -l_{n-2}' \\ -l_{n-1}' \end{pmatrix} \begin{pmatrix} 1 & -u_1' & -u_2' & \cdots & -u_{n-2}' & -u_{n-1}' \end{pmatrix}.$$

 ISBN: 978-84-608-6082-2

Example 1. *Given the bordered tridiagonal matrix*

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

we obtain $b_0' = \dfrac{\det A_n}{\det \widehat{T}_{n-1}} = \dfrac{-133}{11}$ *and* $\widehat{T}_{n-1}^{-1} = \begin{pmatrix} -\frac{1}{2} & \frac{3}{2} & \frac{3}{11} & \frac{9}{22} \\ 0 & 1 & \frac{2}{11} & \frac{3}{11} \\ 0 & 0 & \frac{2}{11} & \frac{3}{11} \\ 0 & 0 & \frac{5}{11} & \frac{2}{11} \end{pmatrix}$. *The UTL factorization of* A_n *is trivial. Also for* $A_n^{-1} = L_n^{-1}T_n^{-1}U_n^{-1}$. *The inverse of a tridiagonal plus a rank-one matrix decomposition for* A_n^{-1} *is*

$$A_n^{-1} = \left(\begin{array}{c|cccc} 0 & & & & \\ \hline & -\frac{1}{2} & \frac{3}{2} & \frac{3}{11} & \frac{9}{22} \\ & 0 & 1 & \frac{2}{11} & \frac{3}{11} \\ & 0 & 0 & \frac{2}{11} & \frac{3}{11} \\ & 0 & 0 & \frac{5}{11} & \frac{2}{11} \end{array}\right) + \frac{-11}{133} \begin{pmatrix} 1 \\ \frac{35}{11} \\ \frac{38}{11} \\ \frac{5}{11} \\ -\frac{15}{11} \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{2} & \frac{11}{2} & \frac{18}{11} & \frac{43}{22} \end{pmatrix}.$$

3 Finite and infinite tridiagonal matrices

A tridiagonal matrix is unreduced if the entries on its subdiagonal and superdiagonal are different to zero. We recall here some well-known lemmas about tridiagonal matrices. The proof can be found in [2,4].

Lemma 1. *Let* T_n *be a nonsingular matrix, then* $T_n = (t_{ij})_{i,j=1}^n$ *is a tridiagonal unreduced matrix if and only if its inverse matrix* $B = (b_{ij})_{i,j=1}^n$ *has the form*

$$b_{ij} = \begin{cases} p_i q_j, & \text{if } \ i \geq j, \\ r_i s_j, & \text{if } \ j \geq i, \end{cases}$$

with $p_i, q_j, r_i, s_j \neq 0$.

Lemma 2. *Let* T *be an infinite invertible matrix. Then* T *is a tridiagonal unreduced matrix* $T = \{a_i, b_i, c_i\}$ *if and only if its classical inverse matrix* $B = (b_{ij})_{i,j=1}^{\infty}$ *has the entries*

$$b_{ij} = \begin{cases} p_i q_j, & \text{if } \ i \geq j, \\ r_i s_j, & \text{if } \ j \geq i. \end{cases}$$

 ISBN: 978-84-608-6082-2

The vectors $P=(p_1,p_2,...)$, $Q=(q_1,q_2,...)$, $R=(r_1,r_2,...)$, $S=(s_1,s_2,...)$ *satisfy the following recurrence relations*

$$\begin{cases} p_2=\dfrac{1-b_0p_1q_1}{c_1q_1} \\ p_i=\dfrac{-a_{i-2}p_{i-2}-b_{i-2}p_{i-1}}{c_{i-1}} \end{cases} \quad \begin{cases} q_2=\dfrac{-b_0q_1}{a_1} \\ q_i=\dfrac{-c_{i-2}q_{i-2}-b_{i-2}q_{i-1}}{a_{i-1}} \end{cases}$$

$$\begin{cases} r_2=\dfrac{-b_0r_1}{c_1} \\ r_i=\dfrac{-a_{i-2}r_{i-2}-b_{i-2}r_{i-1}}{c_{i-1}} \end{cases} \quad \begin{cases} s_2=\dfrac{1-b_0r_1s_1}{a_1r_1} \\ s_i=\dfrac{-a_{i-2}s_{i-2}-b_{i-2}s_{i-1}}{a_{i-1}} \end{cases}$$

for $i\geq 3$, *with* $q_1\neq 0$, $r_1\neq 0$, *and* $p_1q_1=r_1s_1$.

Remark 1. *In the establishment of the preceding lemma, we have recurrence relations and the conditions* $q_1\neq 0$, $r_1\neq 0$, *and* $p_1q_1=r_1s_1$. *Then, three of these parameters are free. When we choose different values of these, we obtain different classical inverses of the infinite tridiagonal matrix* T.

In the case of finite or infinite tridiagonal matrices with nonzero determinants on their main sections, we have two interesting lemmas:

Lemma 3. *Let* T_n *be a nonsingular tridiagonal matrix given by*

$$T_n=\begin{pmatrix} b_0 & c_1 & & & & \\ a_1 & b_1 & c_2 & & & \\ & a_2 & b_2 & c_3 & & \\ & & a_3 & \ddots & \ddots & \\ & & & \ddots & b_{n-2} & c_{n-1} \\ & & & & a_{n-1} & b_{n-1} \end{pmatrix},$$

with the determinants of its main sections different to zero. Then, matrix T_n *has the LDU factorization given by* $T_n=L_nD_nU_n=$

$$\begin{pmatrix} 1 & & & & & \\ w_1 & 1 & & & & \\ & w_2 & 1 & & & \\ & & w_3 & \ddots & & \\ & & & \ddots & 1 & \\ & & & & w_{n-1} & 1 \end{pmatrix}\begin{pmatrix} b'_0 & & & & & \\ & b'_1 & & & & \\ & & b'_2 & & & \\ & & & \ddots & & \\ & & & & b'_{n-2} & \\ & & & & & b'_{n-1} \end{pmatrix}\begin{pmatrix} 1 & v_1 & & & & \\ & 1 & v_2 & & & \\ & & 1 & v_3 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & v_{n-1} \\ & & & & & 1 \end{pmatrix},$$

where the entries b'_i *can be evaluated recursively by* $b'_0=b_0$, $b'_i=b_i-\frac{a_ic_i}{b'_{i-1}}$, *for* $i=1,2,3,...,n-1$. *The entries* $w_i=\frac{a_i}{b'_{i-1}}$, $v_i=\frac{c_i}{b'_{i-1}}$, *for* $i=1,2,...,n-1$. *This factorization is unique.*

 ISBN: 978-84-608-6082-2

The proof is well known. It is based in taking the product of three matrices on the right-hand side and to identify with the matrix T_n. Then the relations of the establishment hold.

If one or more, a_i or c_i are zero, then $w_i = \frac{a_i}{b_{i-1}} = 0$, or $v_i = \frac{c_i}{b_{i-1}} = 0$, and $b'_{i-1} = b_{i-1}$, as we can see in the next example.

Example 2. *The factorization of the next tridiagonal matrix is*

$$\left(\begin{array}{ccc|cc} b_0 & c_1 & & & \\ a_1 & b_1 & c_2 & & \\ & a_2 & b_2 & c_3 & \\ \hline & & 0 & b_3 & c_4 \\ & & & a_4 & b_4 \end{array}\right) = \left(\begin{array}{ccc|cc} 1 & & & & \\ w_1 & 1 & & & \\ & w_2 & 1 & & \\ \hline & & 0 & 1 & \\ & & & w_4 & 1 \end{array}\right)\left(\begin{array}{ccc|cc} b'_0 & & & & \\ & b'_1 & & & \\ & & b'_2 & & \\ \hline & & & b_3 & \\ & & & & b'_4 \end{array}\right)\left(\begin{array}{ccc|cc} 1 & v_1 & & & \\ & 1 & v_2 & & \\ & & 1 & v_3 & \\ \hline & & & 1 & v_4 \\ & & & & 1 \end{array}\right).$$

They are $w_3 = \frac{a_3}{b'_3} = 0$, $b'_3 = b_3$, $w_4 = \frac{a_4}{b_3}$. *It is obvious, by induction, that*

$$\det T_n = \prod_{i=0}^{n-1} b'_i \qquad \text{and} \qquad b'_i = \frac{\det T_{i+1}}{\det T_i} = \frac{\prod_{k=0}^{i} b'_k}{\prod_{i=0}^{i-1} b'_k}.$$

The existence of this factorization is a consequence of the nonsingularity of the main sections of the matrix T_n. We need $4n-2$ flops for such a factorization.

Lemma 4. *Let T be an invertible tridiagonal matrix given by*

$$T = \begin{pmatrix} b_0 & c_1 & & & \\ a_1 & b_1 & c_2 & & \\ & a_2 & b_2 & c_3 & \\ & & a_3 & b_3 & \ddots \\ & & & \ddots & \ddots \end{pmatrix},$$

with $\det T_k \neq 0, k = 1, 2, ...$ *Then, matrix T has the LDU factorization given by*$T = LDU =$

$$= \begin{pmatrix} 1 & & & & \\ w_1 & 1 & & & \\ & w_2 & 1 & & \\ & & w_3 & 1 & \\ & & & \ddots & \ddots \end{pmatrix}\begin{pmatrix} b'_0 & & & & \\ & b'_1 & & & \\ & & b'_2 & & \\ & & & b'_3 & \\ & & & & \ddots \end{pmatrix}\begin{pmatrix} 1 & v_1 & & & \\ & 1 & v_2 & & \\ & & 1 & v_3 & \\ & & & 1 & \ddots \\ & & & & \ddots \end{pmatrix},$$

where entries b'_i can be evaluated recursively by $b'_0 = b_0$, $b'_i = b_i - \dfrac{a_i c_i}{b'_{i-1}}$, for $i = 1, 2, 3, ...$, and $w_i = \dfrac{a_i}{b'_i}$, $v_i = \dfrac{c_i}{b'_i}$, for $i = 1, 2, ...$ This factorization is unique and consistent, because every finite section of T can be factored as a product of main sections of factor matrices.

 ISBN: 978-84-608-6082-2

Proof. Taking the product of the matrices on the right-hand side, we have

$$\begin{pmatrix} b'_0 & b'_0v_1 & & & & \\ b'_0w_1 & b'_1+b'_0v_1w_1 & b'_1v_2 & & & \\ & b'_1w_2 & b'_2+b'_1v_2w_2 & b'_2v_3 & & \\ & & b'_2w_3 & b'_3+b'_2v_3w_3 & \ddots & \\ & & & \ddots & \ddots & \end{pmatrix}$$

that coincides with the matrix T. Therefore, the relations of the establishment hold. □

Consistency of matrices has been studied in [4] for Hessenberg matrices and for tridiagonal matrices. That is, if B is the inverse matrix of a Hessenberg or tridiagonal matrix A, sections of order $k-1$ of A can be calculate deleting the last column and the last row of $([B]_k)^{-1}$ and the two last columns and the two last rows of $([B]_{k+1})^{-1}$. Therefore, this class of infinite unreduced Hessenberg or tridiagonal matrices can be built section to section from B.

We observe that, if one or more, a_i or c_i are null, the $w_i = \frac{a_i}{b_{i-1}} = 0$, or $v_i = \frac{c_i}{b_{i-1}} = 0$, and $b'_{i-1} = b_{i-1}$.

4 An UTL factorization for infinite bordered tridiagonal matrices

We recall that if $A = (a_{ij})_{i,j=1}^{\infty}$ is an infinite matrix of complex numbers, the matrix $B = (b_{ij})_{i,j=1}^{\infty}$ is a *classical inverse* of A if we have $AB = BA = I$. It is well known that an infinite matrix can have not inverse matrix, for example matrix corresponding to right-shift operator, $S_R = (s_{ij})$, with $s_{ij} = 1$ if $i = j+1$, and $s_{ij} = 0$ in other case, has not associated an inverse matrix. It is also well known that an infinite matrix can have two classical inverse matrices, as we can see in [2, 4, 13], and then infinitely many classical inverses, because if B' and B'' are inverses of A, then $\alpha B' + (1-\alpha)B''$ is also an inverse matrix of A, for every $\alpha \in \mathbb{C}$.

The infinite bordered tridiagonal matrices are the form

$$A = \begin{pmatrix} b_0 & u_1 & u_2 & u_3 & u_4 & \cdots \\ l_1 & b_1 & c_2 & & & \\ l_2 & a_2 & b_2 & c_3 & & \\ l_3 & & a_3 & b_3 & c_4 & \\ l_4 & & & a_4 & b_4 & \ddots \\ \vdots & & & & \ddots & \ddots \end{pmatrix}. \tag{2}$$

 ISBN: 978-84-608-6082-2

Theorem 2. *Let A be an infinite invertible bordered tridiagonal matrix as given in (2), and $\widehat{T}$ the infinite invertible tridiagonal matrix obtained from A when the first row and column are deleted. If $\widehat{T}$ is an unreduced matrix, that is, $a_k \neq 0$, $c_k \neq 0$, $k = 2, 3, ...$, and b'_0 is finite and nonzero, then matrix A can be factored in the form UTL as $A = UTL =$*

$$= \begin{pmatrix} 1 & u'_1 & u'_2 & u'_3 & \cdots \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & \ddots \end{pmatrix} \left(\begin{array}{c|cccc} b'_0 & 0 & & & \\ \hline 0 & b_1 & c_2 & & \\ & a_2 & b_2 & c_3 & \\ & & a_3 & b_3 & \ddots \\ & & & \ddots & \ddots \end{array}\right) \begin{pmatrix} 1 & & & & \\ l'_1 & 1 & & & \\ l'_2 & & 1 & & \\ l'_3 & & & 1 & \\ \vdots & & & & \ddots \end{pmatrix},$$

where entries can be calculate recursively, from $u'_1 = \mu$, $l'_1 = \lambda$, and

$$\begin{aligned} u'_k &= \frac{u_k - c_{k-1}u'_{k-2} - b_{k-1}u'_{k-1}}{a_k}, \qquad u'_0 = 0, \\ l'_k &= \frac{l_k - a_{k-1}l'_{k-2} - b_{k-1}l'_{k-1}}{c_k}, \qquad l'_0 = 0, \\ b'_0 &= b_0 - \sum_{i=1}^{\infty} l'_i(c_i u'_{i-1} + b_i u'_i + a_{i+1}u'_{i+1}), \qquad k = 2, 3, ... \end{aligned}$$

Proof. Taking the product UTL and identify with matrix A we obtain

$$\begin{aligned} u_i &= u'_{i-1}c_i + u'_i b_i + u'_{i+1}a_{i+1} \\ l_i &= l'_{i-1}a_i + l'_i b_i + l'_{i+1}c_{i+1}, \qquad i = 1, 2, ..., n-1, \\ b_0 &= b'_0 + \sum_{i=1}^{\infty} l'_i(u'_{i-1}c_i + u'_i b_i + u'_{i+1}a_{i+1}) \end{aligned}$$

with $a_1 = c_1 = 0$. Then, formulas of establishment holds. □

Example 3. *Given the matrix*

$$A = \begin{pmatrix} 4 & 4 & 2 & 1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \cdots \\ 2 & 1 & 4 & & & & & & \\ 1 & 8 & 1 & 2 & & & & & \\ \frac{1}{2} & & 4 & 1 & 1 & & & & \\ \frac{1}{4} & & & 2 & 1 & \frac{1}{2} & & & \\ \frac{1}{8} & & & & 1 & 1 & \frac{1}{4} & & \\ \frac{1}{16} & & & & & \frac{1}{2} & 1 & \frac{1}{8} & \\ \frac{1}{32} & & & & & & \frac{1}{4} & 1 & \ddots \\ \vdots & & & & & & & \ddots & \ddots \end{pmatrix}$$

 ISBN: 978-84-608-6082-2

with the recurrences of Theorem 2, taking $\lambda = \mu = 1$, we obtain an UTL factorization of A,

$$\begin{pmatrix} 1 & 1 & \frac{3}{8} & \frac{-19}{32} & \frac{27}{64} & \frac{43}{64} & \frac{-81}{64} & \cdots \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \end{pmatrix} \begin{pmatrix} 3 & & & & & & & \\ & 1 & 4 & & & & & \\ & 8 & 1 & 2 & & & & \\ & & 4 & 1 & 1 & & & \\ & & & 2 & 1 & \frac{1}{2} & & \\ & & & & 1 & 1 & \frac{1}{4} & \\ & & & & & \frac{1}{2} & 1 & \ddots \\ & & & & & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} 1 & & & & & & & \\ 1 & 1 & & & & & & \\ \frac{1}{4} & & 1 & & & & & \\ \frac{-1}{8} & & & 1 & & & & \\ \frac{9}{8} & & & & 1 & & & \\ \frac{-5}{4} & & & & & 1 & & \\ 1 & & & & & & 1 & \\ \vdots & & & & & & & \ddots \end{pmatrix}$$

where is $b'_0 = 3$. With other values of λ and μ we obtain different UTL factorizations.

Corollary 3. *The inverse of an infinite invertible bordered tridiagonal matrix A, satisfying Theorem 2, has a factorization of the form*

$$A^{-1} = L^{-1}T^{-1}U^{-1}.$$

Such a factorization for A^{-1} is not unique because we can choose parameters λ and μ, and T^{-1} is not unique.

Proof. The product of these matrices is associative, $AA^{-1} = (UTL)(L^{-1}T^{-1}U^{-1}) = I$, and $A^{-1}A = (L^{-1}T^{-1}U^{-1})(UTL) = I$. □

Corollary 4. *Let A be an infinite invertible bordered tridiagonal matrix under the assumptions of Theorem 2. Its classical inverse matrices A^{-1} can be decomposed as a tridiagonal plus a rank-one matrix of the form*

$$A^{-1} = \left(\begin{array}{c|c} 0 & \\ \hline & \widehat{T}^{-1} \end{array}\right) + \frac{1}{b'_0} \begin{pmatrix} 1 \\ -l'_1 \\ -l'_2 \\ -l'_3 \\ \vdots \end{pmatrix} \begin{pmatrix} 1 & -u'_1 & -u'_2 & -u'_3 & \cdots \end{pmatrix}.$$

Different matrices A^{-1} can be obtained because $\widehat{T}^{-1}$ is not unique.

Using Theorem 1 and Lemma 3, for appropriate bordered tridiagonal matrices such that the main sections of $\widehat{T}$ have nonzero determinants, we have a new factorization given by Lemma 5 for finite bordered matrices. Using Theorem 2 and Lemma 4, for appropriate infinite bordered matrices with determinants of main sections of $\widehat{T}$ different to zero, we have Theorem 3.

Lemma 5. *Every nonsingular bordered tridiagonal matrix A_n, as given by (1), with* $\det \widehat{T}_k \neq 0$, $k = 1, 2, ..., n-1$, *has a unique $UL'DU'L$ factorization of the form $A_n = U_n L'_n D_n U'_n L_n$, with*

$$U_n = \begin{pmatrix} 1 & u'_1 & u'_2 & \cdots & u'_{n-1} \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}, L'_n = \begin{pmatrix} 1 & & & & \\ w_1 & 1 & & & \\ & w_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & w_{n-1} & 1 \end{pmatrix}, D_n = \begin{pmatrix} b'_0 & & & \\ & d_1 & & \\ & & \ddots & \\ & & & d_{n-1} \end{pmatrix},$$

$$U'_n = \begin{pmatrix} 1 & v_1 & & & \\ & 1 & v_2 & & \\ & & 1 & \ddots & \\ & & & \ddots & v_{n-1} \\ & & & & 1 \end{pmatrix}, L_n = \begin{pmatrix} 1 & & & & \\ l'_1 & 1 & & & \\ l'_2 & & 1 & & \\ \vdots & & & \ddots & \\ l'_{n-1} & & & & 1 \end{pmatrix},$$

where entries can be calculate recursively in $O(n)$ time.

Theorem 3. *Let A be an infinite invertible bordered tridiagonal matrix as given by (2), let $\widehat{T}$ be the infinite invertible tridiagonal matrix, with the determinants of its main sections different to zero, obtained from A when the first row and column are deleted. We suppose also that matrix $\widehat{T}$ is unreduced and $b'_0 \neq 0$, then matrix A can be factored in the form*

$$\begin{pmatrix} 1 & u'_1 & u'_2 & \cdots \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} 1 & & & \\ w_1 & 1 & & \\ & w_2 & 1 & \\ & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} b'_0 & & & \\ & d_1 & & \\ & & d_2 & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} 1 & v_1 & & \\ & 1 & v_2 & \\ & & 1 & \ddots \\ & & & \ddots \end{pmatrix} \begin{pmatrix} 1 & & & \\ l'_1 & 1 & & \\ l'_2 & & 1 & \\ \vdots & & & \ddots \end{pmatrix}.$$

Corollary 5. *The inverse matrix of A, under the assumptions of Theorem 3, can be evaluated as*

$$A^{-1} = L^{-1}(U')^{-1}D^{-1}(L')^{-1}U^{-1}.$$

We have worked only under the idea of infinite matrices. Under some conditions to be studied, infinite bordered tridiagonal matrices can be regarded as bounded linear operators on ℓ^2. It is an interesting open line.

5 Conclusions

We have proposed a method to factorize finite nonsingular bordered matrices and to obtain their inverse matrices. This method allows us the factorization and the inversion under some conditions of infinite invertible bordered tridiagonal matrices, real or complex. Some examples are also given.

References

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

[2] J. Abderramán Marrero, V. Tomeo, E. Torrano, *Inversion of infinite tridiagonal matrices*, Proceeding of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2014 Vol **1** (2014) 5–16.

[3] J. Abderramán Marrero, V. Tomeo, *Infinite invertible arrowhead matrices and applications*, Proceeding of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2015 Vol **1** (2015) 1–11.

[4] J. Abderramán Marrero, V. Tomeo, E. Torrano, *On inverses of infinite Hessenberg matrices*, J. Comp. Appl. Math. **275** (2015) 356–365.

[5] S. Barnett, *Matrices: Methods and applications*, Oxford University Press, New York, 1990

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

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

[8] C.F. Fischer, R.A. Usmani, *Properties of some tridiagonal matrices and their applications to boundary value problems*, SIAM J. Numer. Anal. **6** (1969) 127–142.

[9] C.M. da Fonseca, J. Petronilho, *Explicit inverses of some tridiagonal matrices*, Linear Algebra and its Aplications **325** (2001) 7–21.

[10] J. Jia, S. Li, *On the inverse and determinant of general bordered tridiagonal matrices*, Computers and Mathematics with Applications **69** (2015) 503–509.

[11] M.E. Kanal, *Parallel algorithm on inversion for adjacent pentadiagonal matrices with MPI*, J. Supercomput. **59** (2012) 1071–1078.

[12] K. H. Rosen, *Discrete Mathematics and its Applications*, McGraw-Hill, New York, 2007.

[13] P. N. Sivakumar, K. C. Shivakumar, *A review of infinite matrices and their applications*, Linear Algebra Appl. **430** (2009) 976–998.

[14] L.H. Thomas, *Elliptic problems in linear difference equations over a network*, Watson Scientific Computing Laboratory Report, Columbia University, 1949.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Quantum entanglement in two-electron systems with emphasis on $(d-1)$-Spherium

I. V. Toranzo[1]

[1] *Departamento de Física Atómica, Molecular y Nuclear,*
Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada,
Granada 18071, Spain

emails: ivtoranzo@ugr.es

Abstract

Up to now the analytical computation of the entanglement has been only restricted to a very few systems of interacting particles (with continuous variables). Here we present a short review of such systems and focus on the analytical calculations of the amount of entanglement exhibited by s-states of the quasi-exactly solvable model known as Spherium. This system is composed of two particles (electrons) interacting via a Coulomb potential and confined to a (d1)- sphere (that is, to the surface of a d-dimensional ball). It is shown the functional dependence of entanglement on the physical parameters that characterize the Spherium: the radius R of the system, the spatial dimensionality d, and the energy, E. The trends found here are then discussed and compared with those observed in the two-electron atomic-like models which will be presented in the beginning.

Key words: quantum entanglement, two-electron models, spherium, quasi-soluble models

1 Introduction

Entanglement, which involves non-separability of degrees of freedom of composite quantum systems, is nowadays regarded as one of the most fundamental phenomena in quantum physics [1, 2, 3]. Entangled states of multipartite quantum systems are endowed with non-classical correlations that give rise to a diverse family of physical phenomena of both fundamental and technological significance. Quantum entanglement can be viewed in two complementary ways. On the one hand, entanglement constitutes a valuable resource. The

 ISBN: 978-84-608-6082-2

controlled manipulation of entangled states is central to several quantum information technologies. On the other hand, entanglement can be regarded as a fundamental ingredient for the physical characterization of natural quantum systems such as, for instance, atoms and molecules. These two points of view are closely related to each other, although the latter is somehow less developed than the former.

Exactly solvable and semi-solvable systems (for which its Schrödinger eigenvalue equation can be solved in a closed analytical fashion for particular eigenstates) provide a valuable arena for the exploration of the entanglement properties of quantum systems of interacting particles. In particular, they provide useful insights for illuminating the entanglement-related features of natural and artificial atomic systems. Unfortunately, there are few such systems where entanglement measures can be evaluated analytically. In fact, up until now, the only system of two interacting particles with continuous variables where entanglement has been calculated in an exact analytical way is the Moshinsky model [4, 5]. Even for the Crandall and Hooke models, entanglement calculations are based upon the numerical evaluation of rather complex multi-dimensional integrals [6].

2 $(d-1)$-Spherium

In the present contribution we show that Spherium, a system of two particles (electrons) interacting through a Coulomb potential and confined to a $(d1)$-sphere (i.e., to the surface of a d-dimensional ball), is a highly exceptional model, where the amount of entanglement exhibited by some of its eigenstates can be determined in an exact and fully analytical way [7].

The entanglement of two-electron systems is computed through what is known as linear entropy, which is defined as

$$\xi\left[|\Phi\rangle\right] \,=\, 1 - 2\,\mathrm{Tr}\left[\left(\rho_1^{(\mathrm{coord.})}\right)^2\right]\,\mathrm{Tr}\left[\left(\rho_1^{(\mathrm{spin})}\right)^2\right], \tag{1}$$

where $\rho_1^{(\mathrm{coord.})}$ denotes the marginal density matrix obtained after computing the partial trace of the matrix density $\rho^{(\mathrm{coord.})}$ over the coordinates of one of the particles

$$\rho_1 = \mathrm{Tr}_{2,3,\ldots,N}\left(|\Phi\rangle\langle\Phi|\right) \tag{2}$$

and $\rho_1^{(\mathrm{spin})}$ denotes the marginal spin density matrix.

The Hamiltonian corresponding to the Spherium, in atomic units, reads

$$H = -\frac{\nabla_1^2}{2} - \frac{\nabla_2^2}{2} + \frac{1}{r_{12}}, \tag{3}$$

 ISBN: 978-84-608-6082-2

where $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the interelectronic distance. The corresponding Schrödinger equation can be cast in the form

$$\left[\frac{u^2}{4R^2} - 1\right]\frac{d^2\Psi}{du^2} + \left[\frac{u(2d-3)}{4R^2} - \frac{d}{u}\right]\frac{d\Psi}{du} + \frac{\Psi}{u} = E\Psi \tag{4}$$

with $u = r_{12}$. Loos and Gill recently proved in [8] that equation (4) admits closed analytical solutions for particular, discrete values of the radius $R = R_{n,m}$. These exact eigenfunctions of the Spherium system have a polynomial form,

$$\Psi_{n,m}(r_{12}) = \sum_{k=0}^{n} s_{k,m} r_{12}^k \tag{5}$$

where the coefficients $s_{k,m} \equiv s_{k,m}(d)$ are determined by the recurrence relation

$$s_{k,m} = \frac{s_{k+1,m} + \left[k(k+2(d-1)-2)\frac{1}{4R_{n,m}^2} - E_{n,m}\right] s_{k,m}}{(k+2)(k+(d-1))}, \tag{6}$$

with the starting values $s_{0,m} = 1$ and $s_{1,m} = \frac{1}{(d-1)-1} \equiv \gamma$. The integer parameter n has values $n = 1, 2, \ldots$ and m is the number of roots that the polynomial (5) has in the range $[0, 2R]$. That is, the wavefunction (5) corresponds to the mth excited s-state. For a given n, the energies are obtained by finding the roots of the equation $s_{n+1,m} = 0$, which is a polynomial in E, of degree $\frac{n+1}{2}$. The corresponding radius $R_{n,m}$ is found through the relation

$$R_{n,m}^2 E_{n,m} = \frac{n}{2}\left(\frac{n}{2} + (d-1) - 1\right). \tag{7}$$

In order to compute the entanglement of the Spheriums eigenstates (with $m = 0$) we work with appropriately normalized eigenfunctions

$$\psi_{n,0} = \frac{\Psi_{n,0}}{R^{d-1} N_n^{1/2}} \tag{8}$$

where $N_n = \int |\Psi_{n,0}|^2 \, d\Omega_1 d\Omega_2$.

The analytical evaluation of the entanglement for the wavefuntions $\psi_{n,0}(r_{12})$ implies the computation of the normalization constant N_n as well as the trace $\mathrm{Tr}\left[\left(\rho_1^{(\mathrm{coord.})}\right)^2\right]$ which is given by the multidimensional integral

$$\begin{aligned}
\mathrm{Tr}\left[\left(\rho_1^{(\mathrm{coord.})}\right)^2\right] &= \int_{\mathbb{R}^{4(d-1)}} \psi_{n,0}(\mathbf{r}_1', \mathbf{r}_2)\psi_{n,0}^*(\mathbf{r}_1', \mathbf{r}_2)\psi_{n,0}^*(\mathbf{r}_1', \mathbf{r}_2')\psi_{n,0}(\mathbf{r}_1, \mathbf{r}_2') \\
&\quad \times R^{4(d-1)} d\Omega_1 d\Omega_1 d\Omega_1' d\Omega_2' \\
&= N_n^{-2}(I_0 + 4\gamma I_1 + 6\gamma^2 I_2 + 4\gamma^3 I_3 + \gamma^4 I_4),
\end{aligned} \tag{9}$$

 ISBN: 978-84-608-6082-2

expressed in terms of the integrals $\{I_i\}_{i=1}^{4}$, whose explicit expressions can be found for a given state in a compact form. This is particularly true for the ground state $\psi_{1,0}(r_{12})$ in which case the entanglement measure, according to (1), is given by

$$\xi[|\psi_{1,0}\rangle] \quad = \quad 1 - \mathrm{Tr}\left[\left(\rho_1^{(\mathrm{coord.})}\right)^2\right]$$

The main entanglement features of the ground state for the Spherium will be discussed. Briefly, it turns out that they are similar to the ones found in certain exactly solvable atomic models as mentioned above.

Acknowledgements

This research was partially supported by the Projects FQM-7276 and FQM-207 of the Junta de Andalucia and the grant FIS2011-24540 of the Ministerio de Economia y Competitividad (Spain).

References

[1] Bengtsson I. and Zyczkowski K., *Geometry of Quantum States: An Introduction to Quantum Entanglement* (Cambridge: Cambridge University Press, 2006).

[2] Amico L., Fazio L., Osterloh A. and Vedral V., Rev. Mod. Phys. 80 517 (2008).

[3] Tichy M., Mintert F. and Buchleitner A., J. Phys. B: At. Mol. Opt. Phys. 44 192001 (2011).

[4] Yáñez R. J., Plastino A. R. and Dehesa J. S., Eur. Phys. J. D 56 141 (2010).

[5] Bouvrie P. A., Majtey A. P., Plastino A. R., Sanchez-Moreno P. and Dehesa J. S., Eur. Phys. J. D 66 1 (2012).

[6] Manzano D., Plastino A. R., Dehesa J. S. and Koga T., J. Phys. A: Math. Theor. 43 275301 (2010).

[7] Toranzo I. V., Plastino A. R., Sanchez-Moreno P. and Dehesa J. S., J. Phys. A 48 475302 (2015).

[8] Loos P. F. and Gill P. M. W., Phys. Rev. Lett. 103 123008 (2009).

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Transition-metal oxide clusters: structural and magnetic properties, infrared spectra and perspectives with applications in catalysis

M. B. Torres[1], A. Aguado[2], F. Aguilera-Granja[3], A. Vega[2] and L. C. Balbas[2]

[1] *Departmento de Matemáticas y Computación, Universidad de Burgos, Burgos, Spain*

[2] *Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, Valladolid, Spain*

[3] *Instituto de Física, Universidad Autónoma de San Luis de Potosí, San Luis de Potosí, México*

emails: begonia@ubu.es, , , ,

Abstract

Transition-metal (TM) oxide clusters have attracted the attention of the scientific community and are, presently, a hot topic for several reasons. From the technological point of view, these systems have been shown to be good models for characterizing reactive sites in heterogeneous oxidation reactions [1] [2]. TM oxide clusters formed by elements of ferromagnetic bulk are also interesting to evaluate the influence of the oxidation on their inherent magnetic properties [3] [4]. Designing a nanoparticle that retains a net magnetic moment in realistic conditions is appealing in many contexts. For instance, CO-oxide nanoparticles are useful in the storage media and biomedical sensors [5]. Moreover, from the fundamental viewpoint, the physics and chemistry of TM oxides are quite rich. Just characterizing the metallic or insulating character of bulk TM oxides is a challenge for the theoretical models which often have to accurately describe the electronic correlations. Understanding the evolution of the electronic and structural properties, strongly interconnected, as size increases, from the simplest TM-O molecule to the bulk limit, is necessary knowledge if we want to achieve the technological goals.

In this context, we have investigated, by means of density functional theoretical calculations, the structural, electronic and magnetic properties of neutral and ionized oxides of the magnetic dimer FeCo as a function of the oxygen content ($FeCoO_n$ and

 ISBN: 978-84-608-6082-2

$FeCoO_n{}^+$ with n=1-6). The binding energy shows saturation for an oxygen content in the range n=4-6. The metal-metal bond weakens with increasing oxygen content. This is reflected in the metal-to-oxygen charge transfer, although not systematically in the magnetic properties because the total spin moment oscillates as a function of n between high-spin states (characterized by parallel magnetic couplings) and low-spin states (characterized by antiparallel couplings). Oxide clusters in the high-spin state retail the same total moment as the bare FeCo dimer because of the direct contribution of oxygen atoms. Upon ionization, the weakening of the metal-metal bond is less marked and the overall magnetic moments decreases because of the increasing tendency toward antiparallel couplings. We calculated the vibrational frequencies and IR intensities of certain isomers of different geometries and spin states (see Figure 1), from which future IR spectroscopy experiments could confirm the structural pattern and, the magnetic state.

On the other hand, it was shown that small bimetallic clusters exhibit high superior catalytic properties [6]. In a previous work [7], we investigated, by means of first-principles calculations the formation and desorption of CO_2 on oxidized doped-gold complexes $MAu_nnO_m^+$ (M= Ti, Fe; n=1, 4-7, m=1-2). Recently, reactions of nickel oxide cluster ions, $Ni_nO_{n+x}^+$ (n = 4-10, x = -1-3), with CO in a He buffer gas were investigated using mass spectrometry [8]. When the cluster ions react at room temperature, a CO molecule tends to attach readily to $Ni_nO_{n+x}^+$ for all of the cluster ions with different stoichiometry, although rate constants of the CO attachment reaction are more or less stoichiometry-dependent. However, CO was found to be released from the cluster ions when the cluster ions were heated up to 523 K after the reaction. This finding is interpreted experimentally, such that the CO molecule that physisorbs weakly to $Ni_nO_{n+x}^+$ at room temperature desorbs into the gas phase by the post heating. Currently, we are investigating, using computational methods, these processes and we present some prospects for catalytic applications of Ni oxide clusters. Relying on the study previously conducted by Rodrigo et al on cobalt oxide clusters [9], we present here preliminary calculations explaining the experiments, obtained by Sakuma [8], in nickel oxide clusters.

Key words: transition metal oxide clusters, electronic and magnetic properties, vibrational frequencies, reactivity of clusters, nanoscale catalysts

 ISBN: 978-84-608-6082-2

Acknowledgements

We acknowledge the support of the Spanish "Ministerio de Ciencia e Innovación" (Grant FIS2014-59279-P)

References

[1] Y. Xie et al , *Oxidation Reactions on Neutral Cobalt Oxide Clusters: Experimental and Theoretical Studies*, Phys. Chem. Chem. Phys. **12** (2010) 15716-15717.

[2] Z.-Ch. et al, *Gas Phase Neutral Binary Oxide Clusters: Distribution, Structure, and Reactivity towards CO*, J. Phys. Chem. Lett. **3** (2012) 2415–2419.

[3] D. Roy et al, *Magnetic Moment and Local Moment Alignment in Anionic and/or Oxidized Fe_n Clusters*, J. Chem. Phys. **132** (2010) 194305.

[4] Y. Wang et al, *Comparative DFT Study of Structure and Magnetism of TM_nO_m (TM = Sc.Mn, n=1-2, m =1-6) Clusters*, Phys. Chem. Chem. Phys. **12** (2010) 2471–2477.

[5] N. Poudyal et al, *Synthesis of Monodisperse FeCo Nanoparticle by Reductive Salt-Matrix Annealing*, Nanotechnology **24** (2013) 345605.

[6] C. Dupont et al, J. Phys. Chem. C **112** (2008) 18062.

[7] M. B. Torres et al, *First Principles Study of CO Adsorption-CO_2 Desorption Mechanisms on Oxidized Doped-Gold Cationic Clusters $MAu_nO_m^+$ (M=Ti, Fe; n = 1, 4-7; m=1-2)*, Int. J. Quantum Chem **111** (2011) 510–519.

[8] K. Sakuma et al, *Oxidation of CO by Nickel Oxide Clusters Revealed by Post Heating*, J. Phys. Chem. A **117** (2013) 3260–3265.

[9] R. Aguilera-del-Toro, F. Aguilera-Granja, A. Vega, L.C. Balbás *Structure, fragmentation patterns, and magnetic properties of small cobalt oxide clusters* Phys. Chem. Chem. Phys. **16** (2014) 21732–21741

 ISBN: 978-84-608-6082-2

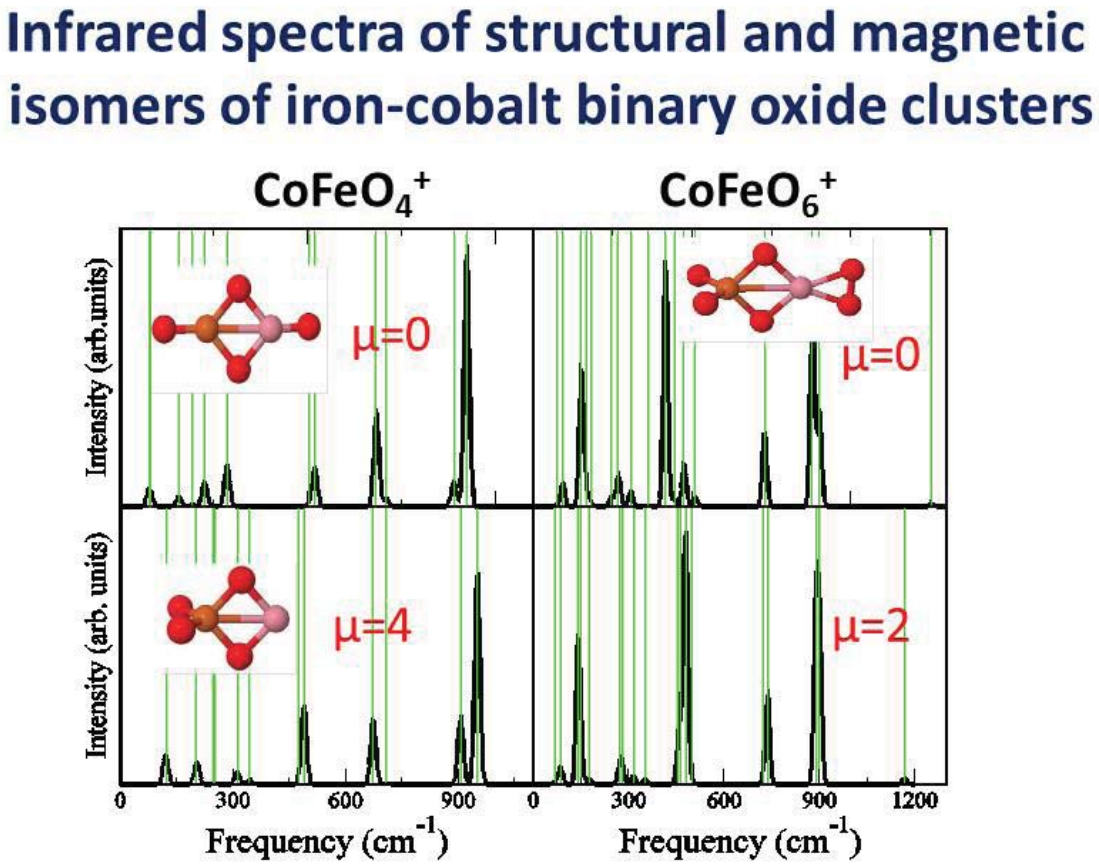


Figure 1: Summary of vibrational spectra for n = 6

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 4–8 July, 2016.

Golden Dual Fullerenes and their Topological Relationship to Fullerenes

Lukas Trombach[1], Sergio Rampino[2], Lai-Sheng Wang[3] and Peter Schwerdtfeger[1]

[1] *Centre for Theoretical Chemistry and Physics, The New Zealand Institute for Advanced Study, Massey University Auckland, Private Bag 102904, 0632 Auckland, New Zealand*

[2] *Istituto di Scienze e Tecnologie Molecolari, Consiglio Nazionale delle Ricerche, c/o Dipartimento di Chimica, Biologia e Biotecnologie, Università degli Studi di Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy*

[3] *Department of Chemistry, Brown University, 324 Brook Street, Providence, Rhode Island 02912, USA*

emails: L.Trombach@massey.ac.nz, srampino@thch.unipg.it, lai-sheng_wang@brown.edu, p.a.schwerdtfeger@massey.ac.nz

Abstract

Golden fullerenes have recently been identified by photoelectron spectra by Bulusu et al. [S. Bulusu, Xi Li, L.-S. Wang, X. C. Zeng, PNAS 103, 8326-8330 (2006)]. These unique triangulations of a sphere are related to fullerene duals having exactly 12 vertices of degree five, and the icosahedral hollow gold cages previously postulated are related to the Goldberg-Coxeter transforms of C_{20} starting from a triangulated surface (hexagonal lattice, dual of a graphene sheet). This also relates topologically the (chiral) gold nanowires observed to the (chiral) carbon nanotubes. In fact, the Mackay icosahedra well known in gold cluster chemistry are related topologically to the dual halma transforms of the smallest possible fullerene C_{20}. The basic building block here is the (111) fcc sheet of bulk gold which is dual to graphene. Because of this interesting one-to-one relationship through Euler's polyhedral formula, there are as many golden fullerene isomers as there are fullerene isomers, with the number of isomers N_{iso} increasing polynomially $\sim \mathcal{O}(N_{iso}^9)$. For the recently observed Au_{16}^-, Au_{17}^-, and Au_{18}^- we present simulated photoelectron spectra including all isomers. We also predict the photoelectron spectrum of Au_{32}^-. The stability of the golden fullerenes is discussed in relation with the more compact structures for the neutral and negatively charged Au_{12}

 ISBN: 978-84-608-6082-2

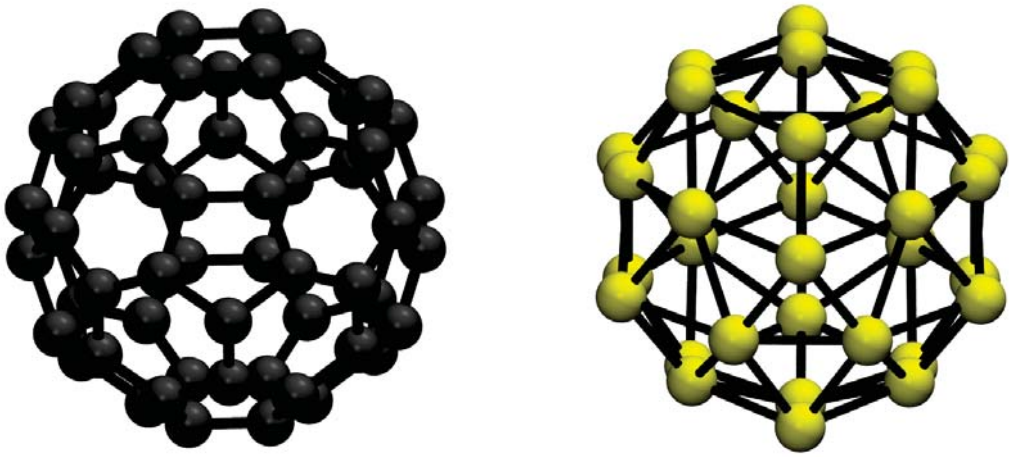


Figure 1: Structure of the icosahedral carbon fullerene C_{60} (left) and the corresponding golden dual fullerene Au_{32}.

to Au_{20} and Au_{32} clusters. As for the compact gold clusters we observe a clear trend in stability of the hollow gold cages towards the (111) fcc sheet. The high stability of the (111) fcc sheet of gold compared to the bulk 3D structure explains the unusual stability of these hollow gold cages.

Key words: Golden dual fullerenes; Topology; Goldberg-Coxeter transforms; Photoelectron spectra; Stability.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Two charges on a plane in a magnetic field: hidden algebra, (particular) integrability, polynomial eigenfunctions

A.V. Turbiner[1] and M A Escobar-Ruiz[2]

[1] *Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México*
Apartado Postal 70-543, 04510 México, D.F., México

[2] *School of Mathematics, University of Minnesota*
Minneapolis, MN 55455, USA

emails: turbiner@nucleares.unam.mx, escob049@umn.edu

Abstract

The quantum mechanics of two Coulomb charges on a plane (e_1, m_1) and (e_2, m_2) subject to a constant magnetic field B perpendicular to the plane is considered. Four integrals of motion are explicitly indicated. For two physically-important particular cases, namely that of two particles of equal Larmor frequencies, $e_c \propto \frac{e_1}{m_1} - \frac{e_2}{m_2} = 0$ (e.g. two electrons) and one of a neutral system (e.g. the electron - positron pair, Hydrogen atom) at rest (the center-of-mass momentum is zero) some outstanding properties occur (i) eigenfunctions are factorizable, they have definite relative angular momentum, (ii) dynamics in radial ρ-direction is the same for both systems, it corresponds to a funnel-type potential and it has hidden $sl(2)$ algebra; at some discrete values of dimensionless magnetic fields $b \leq 1$, (iii) particular integral(s) occur, (iv) the hidden $sl(2)$ algebra emerges in finite-dimensional representation, thus, the system becomes *quasi-exactly-solvable* and (v) a finite number of polynomial eigenfunctions in ρ appear.

Two-body planar Coulomb system, Magnetic field

The quantum mechanics of two non-relativistic Coulomb charges (e_1, m_1) and (e_2, m_2) on a plane subject to a perpendicular constant magnetic field B exhibits many interesting properties both from the point of view of theory and potential applications [1]-[8]. The corresponding Hamiltonian

$$\hat{\mathcal{H}} \;=\; \frac{\left(\hat{\mathbf{p}}_1 - e_1\,\mathbf{A}_{\boldsymbol{\rho}_1}\right)^2}{2\,m_1} + \frac{\left(\hat{\mathbf{p}}_2 - e_2\,\mathbf{A}_{\boldsymbol{\rho}_2}\right)^2}{2\,m_2} + \frac{e_1\,e_2}{\mid \boldsymbol{\rho}_1 - \boldsymbol{\rho}_2 \mid}\;, \tag{1}$$

 ISBN: 978-84-608-6082-2

($\hbar = c = 1$), where $\boldsymbol{\rho}_{1,2}$ and $\hat{\mathbf{p}}_{1,2} = -i\,\nabla_{1,2}$ are the coordinate and momentum of the first (second) particle and $\mathbf{A_r} = \frac{1}{2}\,\mathbf{B} \times \mathbf{r}$ is the vector potential in the symmetric gauge, in general, is not solvable. For arbitrary magnetic field B, the problem possesses three integrals of motion, the total Pseudomomentum [9]

$$\hat{\mathbf{K}} \;=\; \hat{\mathbf{p}}_1 + e_1\,\mathbf{A}_{\boldsymbol{\rho}_1} + \hat{\mathbf{p}}_2 + e_2\,\mathbf{A}_{\boldsymbol{\rho}_2}\;, \tag{2}$$

$[\,\hat{\mathbf{K}},\,\hat{\mathcal{H}}\,] = 0$, and the total angular momentum

$$\hat{\boldsymbol{L}}^{\,\mathrm{T}} \;=\; \boldsymbol{\rho}_1 \times \hat{\mathbf{p}}_1 + \boldsymbol{\rho}_2 \times \hat{\mathbf{p}}_2\;, \tag{3}$$

$[\,\hat{\boldsymbol{L}}^{\,\mathrm{T}},\,\hat{\mathcal{H}}\,] = 0.$

In c.m.s variables

$$\begin{aligned} &\mathbf{R} = \mu_1\,\boldsymbol{\rho}_1 + \mu_2\,\boldsymbol{\rho}_2\;, \quad && \boldsymbol{\rho} = \boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\;, \\ &\hat{\mathbf{P}} \;=\; \hat{\mathbf{p}}_1 + \hat{\mathbf{p}}_2\;, && \hat{\mathbf{p}} \;=\; \mu_2\,\hat{\mathbf{p}}_1 - \mu_1\,\hat{\mathbf{p}}_2\;, \end{aligned} \tag{4}$$

where $\mu_i = \frac{m_i}{M}$ and $M = m_1 + m_2$ is the total mass of the system, $\hat{\mathbf{P}} = -i\,\nabla_{\mathbf{R}},\ \hat{\mathbf{p}} = -i\,\nabla_{\boldsymbol{\rho}}$ are CM and relative momentum, respectively, the Hamiltonian (5) can be gauge transformed into

$$\hat{\mathcal{H}}' \;= U^{-1}\;\hat{\mathcal{H}}\;U \;=\; \frac{\left(\hat{\mathbf{P}} - q\,\mathbf{A}_{\mathbf{R}} - 2\,e_c\,\mathbf{A}_{\boldsymbol{\rho}}\right)^2}{2\,M} + \frac{\left(\hat{\mathbf{p}} - q_{\mathrm{w}}\,\mathbf{A}_{\boldsymbol{\rho}}\right)^2}{2\,m_r} + \frac{e_1\,e_2}{\rho}\;, \tag{5}$$

where $q_{\mathrm{w}} \equiv e_1\,\mu_2^2 + e_2\,\mu_1^2$ is an effective charge (weighted total charge),

$$e_c = (e_1\,\mu_2 - e_2\,\mu_1) \;=\; m_r\left(\frac{e_1}{m_1} - \frac{e_2}{m_2}\right)\;,$$

is a coupling charge,

$$q \;=\; e_1 + e_2\;,$$

the total charge of the system, and

$$U = e^{-i\,e_c\,\mathbf{A}_{\boldsymbol{\rho}}\cdot\mathbf{R}}\;,$$

is a unitary transformation. In two physically relevant cases:

(i) $e_c = 0$, where separation of c.m.s. variables occurs in the Hamiltonian (5) ,

(ii) $q = 0$, for which components of the Pseudomomentum $\hat{\mathbf{K}}$ become commutative ,

the problem becomes superintegrable (the number of integrals is larger than the dimension

 ISBN: 978-84-608-6082-2

of the configuration space) and quasi-exactly-solvable for some discrete values of magnetic field B [10]-[12]. The eigenfunctions Ψ

$$\hat{\mathcal{H}}'\,\Psi \;=\; E\,\Psi\;, \tag{6}$$

$$\Psi \;=\; \chi(\mathbf{R})\,\psi(\boldsymbol{\rho})\;, \tag{7}$$

are factorized in c.m.s. variables. In polar coordinates $\boldsymbol{\rho}=(\rho,\varphi)$, the factor ψ (7) corresponding to the relative motion is factorizable as well

$$\begin{aligned}\psi(\boldsymbol{\rho}) \;&=\; \mathrm{e}^{-a\,B\,\rho^2}\rho^{|s|}\,p(\rho)\,\Phi(\varphi)\,,\\ \Phi(\varphi) \;&=\; \mathrm{e}^{i\,s\,\varphi}\;,\end{aligned} \tag{8}$$

where $\Phi(\varphi)$ is the eigenfunction of $\hat{\ell}_z=-i\,\partial_\varphi$ and consequently $s=0,\,\pm1,\,\pm2,...$ is the magnetic quantum number of the relative motion and $a=a(e_1,m_1,e_2,m_2)>0$ is determined by the parameters of the system. Surprisingly, in both cases $e_c=0$ and $q=0$ the function $p(\rho)$ obeys

$$\Big[-\hat{J}^0_n\hat{J}^- + \hat{J}^+_n - (1+2|s|+\frac{n}{2})\hat{J}^-\Big]p \;=\; \lambda\,p\;, \tag{9}$$

where the first-order differential operators

$$\begin{aligned}&\hat{J}^+_n=\rho^2\partial_\rho - n\,\rho\;, \qquad \partial_\rho\equiv\frac{d}{d\rho}\;,\\ &\hat{J}^0_n=\rho\,\partial_\rho-\frac{n}{2}\,,\\ &\hat{J}^-=\partial_\rho\;,\end{aligned} \tag{10}$$

satisfies the algebra sl_2, n is the spin of representation

$$n \;=\; n(e_1,m_1,e_2,m_2,E)\;,$$

which depends on the physical parameters of the system and the energy E (6), while λ

$$\lambda^2\equiv\frac{B_0}{B}\equiv\frac{1}{b}\;,\quad B_0=B_0(e_1,m_1,e_2,m_2)\;, \tag{11}$$

plays the role of the new spectral parameter.

A hidden algebraic structure occurs [13] (the underlying idea behind quasi-exactly solvability) at nonnegative integer n: the hidden sl_2 algebra appears in finite-dimensional representation and the problem (9) possesses $(n+1)$ eigenfunctions $p_{n,i},\ i=1,\dots(n+1)$ in the form of a polynomial of the nth power. It implies the explicit (algebraic) quantization of the dimensionless parameter λ or equivalently quantization of magnetic field B.

 ISBN: 978-84-608-6082-2

For a given system with fixed masses (m_1, m_2) and charges (e_1, e_2) there exists an infinite, discrete set of values of the magnetic field $B = \{B_n\}$ for which the exact, analytic solutions of the Schrödinger equation occur. The parameter λ takes $[\frac{n+1}{2}]$ positive values and $[\frac{n+1}{2}]$ symmetric negative values, and zero value for even n, for which the problem (9) possesses polynomial solutions, $[z]$ means integer part of z. For even n, always there is a solution at $\lambda_n = 0$. It corresponds to either non-normalizable wavefunction, $a = 0$, or vanishing Coulomb interaction. The interesting fact is that for given n the $[\frac{n+1}{2}]$ physically admitted functions p_n have a number of nodes at $\rho > 0$ varying from zero up to $[\frac{n-1}{2}]$.

More outstanding properties take place. The operator

$$i_n(\rho) \;=\; \prod_{j=0}^{n} (\rho\,\partial_\rho + j) \;, \tag{12}$$

constructed from the Euler-Cartan operator

$$i_n^0 \;=\; \rho\,\partial_\rho - n \;,$$

has a property of annihilator

$$i_n(\rho) : \; \mathcal{P}_n \;\mapsto\; \{0\} \;,$$

where $\mathcal{P}_n \;=\; \langle 1, \rho, \rho^2, ..., \rho^n \rangle$ is the linear space of polynomials in ρ of degree not higher than n. Then

$$[T(n), i_n(\rho)] : \mathcal{P}_n \;\mapsto\; \{0\} \;,$$

and $i_n(\rho)$ is a particular integral [14] with $\mathcal{P}_n$ as the invariant subspace. The eigenfunctions $p_{n,k}$, $k = 1, \ldots (n+1)$ of (9) are the zero modes of $i_n(\rho)$. For the case $e_c = 0$, the gauge rotated operator $i_n(\rho)$

$$\zeta^{(0)}\; i_n(\rho)\; (\zeta^{(0)})^{-1} \;=\; \prod_{j=0}^{n} (\rho\,\mathcal{D}_\rho + j) \equiv \mathcal{I}_n\;(\rho) \;, \tag{13}$$

where

$$\zeta^{(0)} = \mathrm{e}^{-\frac{m_r\,\omega_c\,\rho^2}{4}} \rho^{|s|} \;, \qquad \omega_c = \frac{e_1\,B}{m_1} = \frac{q\,B}{M} \;,$$

and

$$\mathcal{D}_\rho = \partial_\rho + \frac{m_r\,\omega_c\,\rho}{2} - \frac{|s|}{\rho} \;,$$

is the covariant derivative, is a particular integral,

$$[\hat{\mathcal{H}} \;,\; \mathcal{I}_n(\rho)] : \mathcal{V}_n \;\mapsto\; \{0\} \;,$$

at $b = b_n$, or equivalently at $B = B_n$. Hence, for special values of magnetic field the operator $\mathcal{I}_n(\rho)$ is a particular integral with $\mathcal{V}_n \;=\; \zeta^{(0)}\mathcal{P}_n$ as the invariant subspace. In the classical limit the particular integral $\mathcal{I}_n(\rho)$ becomes its classical counterpart $I_n(i\,\rho\,p_\rho)$ [15]. The latter becomes the constant of motion on certain special periodic circular trajectories.

 ISBN: 978-84-608-6082-2

References

[1] M. Robnik, V. G. Romanovski,
Two-dimensional hydrogen atom in a strong magnetic field,
J. Phys. A: Math. Gen. **36** 7923 (2003)

[2] L. A. Burkova, I.E. Dzyaloshinskii, G. F. Drukarev, and B. S. Monozon ,
Hydrogen-like system in crossed electric and magnetic fields,
Sov. Phys. JETP **44**, 276 (1976)

[3] Vincke M. and Baye D.,
Hydrogen molecular ion in an aligned strong magnetic field by the Lagrange-mesh method,
J. Phys. B: At. Mol. Opt. Phys. **39** (2006) 2605 - 18

[4] P. Schmelcher and L. S. Cederbaum,
2-body effects of the Hydrogen-atom in crossed electric and magnetic fields,
Chem. Phys. Lett., 208:548 (1993)

[5] M. Dineykhan, R.G. Nazmitdinov,
Two-electron quantum dot in magnetic field: Analytical Results,
Phys. Rev. B **55**, 13707 (1997)

[6] P.A. Maksym, Tapash Chakraborty,
Quantum Dots in a Magnetic Field: Role of Electron-Electron Interactions,
Phys. Rev. Let. **65** 108 (1990)

[7] A. Soylu, O. Bayrak and I. Boztosun
The Energy Eigenvalues of the Two Dimensional Hydrogen Atom in a Magnetic Field,
Int. J. Mod. Phys. E **15**, 1263 (2006)

[8] Yu. E. Lozovik, I. V. Ovchinnikov, S. Yu. Volkov, L. V. Butov, and D. S. Chemla,
Quasi-two-dimensional excitons in finite magnetic fields,
Phys. Rev. B **65**, 235304 (2002)

[9] L.P. Gorkov, I.E. Dzyaloshinskii,
Contribution to the Theory of the Mott Exciton in a Strong Magnetic Field,
ZhETF **53** (1967) 717-722 *Sov. Phys. JETP* **26** (1968) 449-451

[10] A.V. Turbiner and M.A. Escobar-Ruiz,
Two charges on a plane in a magnetic field: hidden algebra, (particular) integrability, polynomial eigenfunctions,
J. Phys. A **46** , 295204 (2013)

 ISBN: 978-84-608-6082-2

[11] M. Taut,
Two electrons in a homogeneous magnetic field: particular analytical solutions,
J. Phys. A **27**, 1045 (1994)

[12] M. Taut,
Two particles with opposite charge in a homogeneous magnetic field: particular analytical solutions of the two-dimensional Schrödinger equation,
J. Phys. A **32** (1999) 5509 - 5515

[13] A.V. Turbiner,
Quasi-Exactly-Solvable Problems and the $SL(2,R)$ Group,
Comm.Math.Phys. **118**, 467-474 (1988)

[14] A.V. Turbiner,
Particular Integrability and (Quasi)-exact-solvability
J. Phys. A **45** (2013) 025203 (9pp)
`math-ph arXiv:1206.2907`

[15] M.A. Escobar-Ruiz and A.V. Turbiner,
Two charges on a plane in a magnetic field: special trajectories
Journal of Math Physics **54**, 022901 (2013)

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Application of generalized finite difference method to reflection and transmission problems in seismic SH waves propagation

M. Ureña[1], J.J. Benito[1], F. Ureña[2], E. Salete[1], L. Gavete[3] and A. García[1]

[1] *Departamento de Construcción y Fabricación, Universidad Nacional de Educación a Distancia (UNED)*

[2] *Departamento de Matemática Aplicada, Universidad de Castilla-La Mancha (UCLM)*

[3] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid (UPM)*

emails: miguelurenya@gmail.com, jbenito@ind.uned.es, fuprieto@terra.es, esalete@ind.uned.es, lu.gavete@upm.es, angelochurri@gmail.com

Abstract

A matrix formulation of the generalized finite difference method is introduced and a necessary and sufficient condition for the uniqueness of the solution is demonstrated. The method is applied to seismic waves propagation problems, specifically to the problem of reflection and transmission of plane waves in heterogeneous media. Generalized finite differences scheme for SH wave is obtained and the stability of the scheme is analysed. Heterogeneous approximation without nodes at the interface is chosen to solve the problem in heterogeneous media.

Key words: Generalized finite difference method, Seismic wave, Reflection and transmission
MSC 2000: 65M06, 65M12

1 Introduction

Generalized finite differences method (GFDM) is a numerical method based on moving least squares. This method is not dependent on a mesh, reason why it is contained in the well-known group of meshless methods [1]. Given a differential equation, the aim of the

method is to approximate the values of the unknown function at a set of points (nodes). By means of a Taylor expansion, the values of the derivatives at each node are linearised and, if the differential equation is linear, the problem is reduced to solve a linear system of equations. Liska and Orkisz [2] established a current version of the method and Benito et. al. [3] provided explicit formulae.

Finite differences methods are the most widely used in order to solve the seismic wave propagation problem and so Ureña et. al. [4] have incorporated the GFDM into these type of problems. They have studied the stability and the star dispersion in the decoupled system formed by P and SV waves [5] and, moreover, they have obtained schemes for absorbing boundary, using perfectly matched layers (PML) and analysed the influence of several parameters of such schemes [6].

In seismic problems it is usual to consider the soil as a stack of homogeneous layers [7]. In this way, two types of approximation has been considered, homogeneous and heterogeneous approximation. The homogeneous approximation consists in solving the same scheme at each homogeneous medium independently and imposing boundary conditions at the interfaces, see [8] for an example, while the heterogeneous approximation consists in solving the same scheme in all the domain, considering the values of the material properties at each node. In this last case, some authors use an average of the parameters of the surrounding media when nodes are placed at the interface [9].

The paper is organised as follows. Section 1 is an introduction. In section 2 a matrix formulation is given and a necessary and sufficient condition for the uniqueness of the solution is proved. In section 3 the scheme in GFD for SH waves and its stability is obtained. Reflection and transmission is analysed for a heterogeneous approximation without nodes at the interface in section 4 and, finally, conclusions are presented in section 5.

2 GFD method

Definition 1 *Let $D \subset \mathbb{R}^2$ be a domain, M a discretization of D and $\boldsymbol{x}_0 \in M \cap int(D)$. A p-star, or simply a star, with central node $\boldsymbol{x}_0$, is defined as the set $E(\boldsymbol{x}_0) = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\} \subset M$. The p elements of $E(\boldsymbol{x}_0) - \{\boldsymbol{x}_0\}$ are named nodes of the star.*

Let $D \subset \mathbb{R}^2$ be a domain, $\mathbf{x} = (x, y) \in D$ and $f, g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ continuous functions in D. Let $U \in \mathcal{C}^2(\mathbb{R}^2)$ and the problem govern by the second order linear partial differential equation

$$\mathcal{L}_2\big(U(\mathbf{x})\big) = f(\mathbf{x}) \qquad \forall \mathbf{x} \in \Omega = int(D) \tag{1}$$

 ISBN: 978-84-608-6082-2

with boundary condition

$$\mathcal{L}_1\big(U(\mathbf{x})\big) = g(\mathbf{x}) \qquad \forall \mathbf{x} \in \Gamma = fr(D) \tag{2}$$

Let M be a discretization of D and $E(\mathbf{x}_0) = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N\}$ a star. For each $i \in \{1, 2, \ldots, N\}$, $U(\mathbf{x}_0)$ and $U(\mathbf{x}_i)$ are denoted as U_0 and U_i, and the Taylor series centred at $\mathbf{x}_0$ is

$$U_i = U_0 + (\mathbf{x}_i - \mathbf{x}_0)\nabla U_0 + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_0)^T \mathcal{H} U_0 (\mathbf{x}_i - \mathbf{x}_0) + \ldots \tag{3}$$

The terms above second order are ignored and an approximation of second order for U_i is denoted by u_i. The summation of the equations is done

$$\sum_{i=1}^{N} u_i = \sum_{i=1}^{N} \left(u_0 + (\mathbf{x}_i - \mathbf{x}_0)\nabla u_0 + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_0)^T \mathcal{H} u_0 (\mathbf{x}_i - \mathbf{x}_0) \right) \tag{4}$$

The vector of unknowns is denoted by $\mathbf{D}_u$ and the function $B(\mathbf{D}_u)$ is defined as

$$B(\mathbf{D}_u) = \sum_{i=1}^{N} \left(u_0 - u_i + \boldsymbol{\varepsilon}_i^T \mathbf{D}_u \right)^2 w_i^2 \tag{5}$$

where $\mathbf{D}_u = \begin{pmatrix} \dfrac{\partial u_0}{\partial x} & \dfrac{\partial u_0}{\partial y} & \dfrac{\partial^2 u_0}{\partial x^2} & \dfrac{\partial^2 u_0}{\partial x \partial y} & \dfrac{\partial^2 u_0}{\partial y^2} \end{pmatrix}^T$ is the partial derivatives vector, $w_i = w(\mathbf{x}_i)$ is the weighting function and $\boldsymbol{\varepsilon}_i = \begin{pmatrix} h_i & k_i & \dfrac{h_i^2}{2} & h_i k_i & \dfrac{k_i^2}{2} \end{pmatrix}^T$ with $h_i = x_i - x_0$ and $k_i = y_i - y_0$ the relative coordinates to the central node of the star.

The function is expanded and its gradient is calculated (see [10] for matrix derivation)

$$B(\mathbf{D}_u) = \sum_{i=1}^{N} \left[w_i^2 (u_0 - u_i)^2 + w_i^2 (\boldsymbol{\varepsilon}_i^T \mathbf{D}_u)^2 + 2 w_i^2 (u_0 - u_i) \boldsymbol{\varepsilon}_i^T \mathbf{D}_u \right] \tag{6}$$

$$\nabla B(\mathbf{D}_u) = \sum_{i=1}^{N} \left[2 w_i^2 \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \mathbf{D}_u + 2 w_i^2 (u_0 - u_i) \boldsymbol{\varepsilon}_i \right] \tag{7}$$

Making $\nabla B(\mathbf{D}_u) = \mathbf{0}$ in order to minimise the function,

$$\sum_{i=1}^{N} w_i^2 \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \mathbf{D}_u = \sum_{i=1}^{N} w_i^2 (u_i - u_0) \boldsymbol{\varepsilon}_i \tag{8}$$

a system of linear equations is obtained

$$A \mathbf{D}_u = \mathbf{b} \tag{9}$$

where

$$A = \sum_{i=1}^{N} w_i^2 \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \tag{10}$$

$$\mathbf{b} = \sum_{i=1}^{N} w_i^2 (u_i - u_0) \boldsymbol{\varepsilon}_i \tag{11}$$

Solving the system, the values of the partial derivatives are obtained

$$\mathbf{D}_u = A^{-1}\mathbf{b} = A^{-1} \sum_{i=1}^{N} w_i^2 \boldsymbol{\varepsilon}_i u_i - A^{-1} \sum_{i=1}^{N} w_i^2 \boldsymbol{\varepsilon}_i u_0 = \mathbf{D}_u = -\mathbf{m}_0 u_0 + \sum_{i=1}^{N} \mathbf{m}_i u_i \tag{12}$$

with

$$\mathbf{m}_0 = \sum_{i=1}^{N} A^{-1} w_i^2 \boldsymbol{\varepsilon}_i \tag{13}$$

$$\mathbf{m}_i = A^{-1} w_i^2 \boldsymbol{\varepsilon}_i \tag{14}$$

and, as convention, the coefficient of the linearization of the partial derivatives may be represented by $\mathbf{m}_i = \begin{pmatrix} m_{ix} & m_{iy} & m_{ixx} & m_{ixy} & m_{iyy} \end{pmatrix}^T$, $\forall i \in \{0, 1, 2, \ldots, N\}$

Let $\mathbf{a}$ be the coefficient vector in (1) and so $\mathbf{a}^T \mathbf{D}_u = f(\mathbf{x})$. From the value of $\mathbf{D}_u$ in (12), the following expression is obtained

$$-\mathbf{a}^T \mathbf{m}_0 u_0 + \sum_{i=1}^{N} \mathbf{a}^T \mathbf{m}_i u_i = f(\mathbf{x}) \tag{15}$$

Making $\lambda_0 = \mathbf{a}^T \mathbf{m}_0$ and $\lambda_i = \mathbf{a}^T \mathbf{m}_i$ the *equation of the star* is achieved

$$-\lambda_0 u_0 + \sum_{i=1}^{N} \lambda_i u_i = f(\mathbf{x}) \tag{16}$$

Proposition 1 *A is a positive semidefinite matrix.*

Proof: Indeed, $\forall \mathbf{v} \in \mathbb{R}^5 - \{\mathbf{0}\}$,

$$\mathbf{v}^T A \mathbf{v} = \mathbf{v}^T \sum_{i=1}^{N} w_i^2 \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \mathbf{v} = \sum_{i=1}^{N} w_i^2 \mathbf{v}^T \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \mathbf{v} = \sum_{i=1}^{N} w_i^2 (\mathbf{v}^T \boldsymbol{\varepsilon}_i)^2 \geq 0$$

Proposition 2 *The minimum number of nodes in each star, in order to the system $A\boldsymbol{D}_u = \boldsymbol{b}$ has an unique solution, is 5.*

Proof: The system $A\mathbf{D}_u = \mathbf{b}$ has an unique solution when A is a positive definite matrix, meaning, if $\nexists \mathbf{v} \in \mathbb{R}^5 - \{\mathbf{0}\}$ such that $\mathbf{v}^T \boldsymbol{\varepsilon}_i = 0$, $\forall i = 1, 2, \ldots, N$. If N was less than 5, there would be $\mathbf{v} \in \langle \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_N \rangle^{\perp}$ such that $\mathbf{v}^T A \mathbf{v} = \mathbf{0}$. Therefore, the minimum number of nodes in each star is 5.

Corollary 1 *The system $A\boldsymbol{D}_u = \boldsymbol{b}$ has an unique solution if and only if the set $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_N\}$ contains a base of $\mathbb{R}^5$*

Proof: Immediate from above

3 GFDM and SH waves

In seismic wave propagation problems the x-z plane is typically considered perpendicular to Earth's surface and contains a seismic source and a receiver. In this geometry P and SV waves are decoupled from SH waves and can be treated separately [7]. If $V(x, z, t)$ is the y component of the displacement, the equation for a perfectly elastic, homogeneous and isotropic medium is

$$V_{tt} = \beta^2 (V_{xx} + V_{zz}) \tag{17}$$

where $\beta = \sqrt{\frac{\mu}{\rho}}$ is the velocity of the wave and μ and ρ are the shear modulus and the density, respectively.

3.1 GFD scheme

Classical finite differences formula approximates the temporal second order partial derivative. Generalized finite differences formula approximates the spatial second order partial derivatives.

$$v_{0,tt}^n = \frac{v_0^{n-1} - 2v_0^n + v_0^{n+1}}{\triangle t^2} \tag{18}$$

$$v_{0,ii}^n = -m_{0ii} v_0^n + \sum_{k=1}^{N} m_{kii} v_k^n \tag{19}$$

where N is the number of nodes of the star and, for each $i = x, z$, m_{0ii} is the coefficient of the central node of the star and m_{kii} are the coefficient of the rest of nodes of the star.

Substituting the above formulae in (17), the following expression is obtained

$$\frac{v_0^{n-1} - 2v_0^n + v_0^{n+1}}{\triangle t^2} = \beta^2 \Bigg(- m_{0xx} v_0^n + \sum_{k=1}^{N} m_{ixx} v_i^n - m_{0zz} v_0^n + \sum_{k=1}^{N} m_{izz} v_i^n \Bigg) \tag{20}$$

and, solving for the displacement at time $t = (n+1)\triangle t$, the GFD scheme for SH waves is achieved,

$$v_0^{n+1} = \Big[2 - \triangle t^2 \beta^2 (m_{0xx} + m_{0zz}) \Big] v_0^n - v_0^{n-1} + \triangle t^2 \beta^2 \sum_{k=1}^{N} (m_{ixx} + m_{izz}) v_i^n \tag{21}$$

Neumann boundary condition is considered in the case of free surface for which the condition is the traction vector is zero. For the SH wave case, in this geometry, the condition is tangential stress is zero, $\tau_{yz} = 0$. In order to solve this, a new node is added for each node at the Neumann boundary and an equation is obtained for each node at the Neumann boundary. If v_0 is a node at the free surface, then it is followed

$$\tau_{yz} = 0 \implies \mu v_{0,z} = 0 \implies -m_{0z} v_0 + \sum_{i=1}^{N} m_{iz} v_i = 0 \tag{22}$$

If J is the index set for the new nodes and K is the index set for the rest of the nodes, then the equation for v_0 is

$$\sum_{j \in J} m_{jz} v_j = m_{0z} v_0 - \sum_{k \in K} m_{kz} v_k \tag{23}$$

In this way, an auxiliary system of equations is obtained.

3.2 Stability

In order to check the stability of the scheme, the von Neumann method is applied [11]. Therefore, considering the harmonic wave $v = A \exp(i(\omega t - \mathbf{kx}))$ where A is the amplitude, ω the angular frequency, $\mathbf{k} = (k_x, k_z)$ the wavenumber vector and $\mathbf{x} = (x, z) \in M$,

$$v_0^n = A \exp(i\omega n \triangle t) \cdot \exp(-i\mathbf{k}\mathbf{x}_0) = A\xi^n \exp(-i\mathbf{k}\mathbf{x}_0) \tag{24}$$

$$v_j^n = A \exp(i\omega n \triangle t) \cdot \exp(-i\mathbf{k}\mathbf{x}_j) = A\xi^n \exp(-i\mathbf{k}\mathbf{x}_j) \tag{25}$$

where $\xi = \exp(i\omega \triangle t)$, $\mathbf{x}_0 = (x_0, z_0)$ and $\mathbf{x}_j = (x_j, z_j)$

Let's calculate the values of $\triangle t$ satisfying $|\xi| \leq 1$. Substituting (24) and (25) in (21)

$$A\xi^{n+1}\exp(-i\mathbf{k}\mathbf{x}_0) = [2-\beta^2\triangle t^2(m_{0xx}+m_{0zz})]A\xi^n\exp(-i\mathbf{k}\mathbf{x}_0) - \\ -A\xi^{n-1}\exp(-i\mathbf{k}\mathbf{x}_0) + \beta^2\triangle t^2 A\xi^n\sum_{j=1}^{N}(m_{jxx}+m_{jzz})\exp(-i\mathbf{k}\mathbf{x}_j)$$

Keeping in mind that $\mathbf{h}_j = (h_j, l_j) = (x_j - x_0, z_j - z_0)$ and dividing the above equation by $A\xi^{n-1}\exp(-i\mathbf{k}\mathbf{x}_0)$

$$\xi^2 = [2-\beta^2\triangle t^2(m_{0xx}+m_{0zz})]\xi - 1 + \xi\beta^2\triangle t^2\sum_{j=1}^{N}(m_{jxx}+m_{jzz})\exp(-i\mathbf{k}\mathbf{h}_j) \implies \quad (26)$$

$$\implies \xi^2 - \left[2-\beta^2\triangle t^2(m_{0xx}+m_{0zz}) + \beta^2\triangle t^2\sum_{j=1}^{N}(m_{jxx}+m_{jzz})\exp(-i\mathbf{k}\mathbf{h}_j)\right]\xi + 1 = 0 \quad (27)$$

Since $\quad m_{0xx} = \sum_{j=1}^{N} m_{jxx} \quad$ and $\quad m_{0zz} = \sum_{j=1}^{N} m_{jzz}$,

$$\xi^2 - b\xi + 1 = 0 \quad (28)$$

being $b = 2 - \beta^2\triangle t^2\sum_{j=1}^{N}(m_{jxx}+m_{jzz})(1-\exp(-i\mathbf{k}\mathbf{h}_j))$

Applying Cardano-Vieta formulae, solutions in (28) satisfy

$$\begin{cases}\xi_1 + \xi_2 = b \\ \xi_1 \cdot \xi_2 = 1\end{cases} \quad (29)$$

Second equation implies $\xi_2 = \xi_1^{-1}$ and, taking in account that $|\xi_1| \leq 1$ and $|\xi_2| \leq 1$, $|\xi_1| = |\xi_2| = 1$. Now, from the first equation is followed

$$|b| = |\xi_1 + \xi_1^{-1}| = |2\cos\omega\triangle t| \leq 2 \quad (30)$$

From Euler's formula, $1 - \exp(-i\mathbf{k}\mathbf{h}_j) = 1 - \cos(\mathbf{k}\mathbf{h}_j) + i\sin(\mathbf{k}\mathbf{h}_j)$, and so

$$b = 2 - \beta^2\triangle t^2\sum_{j=1}^{N}(m_{jxx}+m_{jzz})(1-\cos(\mathbf{k}\mathbf{h}_j)) - i\cdot\beta^2\triangle t^2\sum_{j=1}^{N}(m_{jxx}+m_{jzz})\sin(\mathbf{k}\mathbf{h}_j)$$

Denoting $p = \sum_{j=1}^{N}(m_{jxx}+m_{jzz})(1-\cos(\mathbf{k}\mathbf{h}_j))$ and $q = \sum_{j=1}^{N}(m_{jxx}+m_{jzz})\sin(\mathbf{k}\mathbf{h}_j)$, the equation is written as

$$b = 2 - \beta^2\triangle t^2 p + i\cdot\beta^2\triangle t^2 q \quad (31)$$

$$|b|^2 = \left(2-\beta^2\triangle t^2 p\right)^2 + \left(\beta^2\triangle t^2 q\right)^2 = \beta^4(p^2+q^2)\triangle t^4 - 4\beta^2 p\triangle t^2 + 4 \quad (32)$$

and, since $|b| \leq 2$, $|b|^2 \leq 4$ and solutions have to satisfy the inequality

$$\beta^4(p^2+q^2)\triangle t^4 - 4\beta^2 p \triangle t^2 \leq 0 \tag{33}$$

This inequality is a biquadratic inequation with positive solutions if, and only if, $p > 0$ in which case it is satisfied

$$0 \leq \triangle t \leq \sqrt{\frac{4p}{\beta^2(p^2+q^2)}} \tag{34}$$

Applying conservative criteria, the scheme is stable for the star considered when

$$0 \leq \triangle t \leq \frac{\sqrt{2}}{\beta} \cdot \sqrt{\frac{2-\sqrt{2}}{m_{0xx}+m_{0zz}}} \leq \sqrt{\frac{4p}{\beta^2(p^2+q^2)}} \tag{35}$$

Therefore, the GFD scheme for the SH wave is stable if the expression is satisfied for the star with the highest sum of m_0 coefficients

$$0 \leq \triangle t \leq \frac{\sqrt{2}}{\beta} \cdot \sqrt{\frac{2-\sqrt{2}}{\left(m_{0xx}+m_{0zz}\right)_{max}}} \tag{36}$$

4 Reflection and transmission

When an incident SH plane wave arrives at an interface of two media, a part is reflected and another part is transmitted. Snell's law provides the angles of reflection and transmission and the imposition of boundary conditions at the interface provides the reflection and transmission coefficients [12]. If the second medium is air, then it is considered as free surface and only reflected waves exit.

In the next cases, the same Ricker pulse is considered

$$U(t) = A(1-2\pi^2 f^2(t-0.05)^2)exp(-\pi^2 f^2(t-0.05)^2), \quad 0 \leq t \leq \frac{\sqrt{2}}{\pi f} \tag{37}$$

where $f = \dfrac{10\sqrt{2}}{\pi}$ Hz is the frequency and $A = 0.5$ m is the amplitude.

4.1 Free surface

The domain $D = [0, 2.1] \times [-1, 0] \subset \mathbb{R}^2$ is considered and a discretization of 37 654 nodes is performed. The medium has $\lambda = 10$ Pa as first Lamé parameter, a shear modulus of $\mu = 4$ Pa and a density of $\rho = 1\ kg/m^3$ and so, the medium has a shear velocity of $\beta = 2\ m/s$. The SH plane wave (37) arrives from the lower right corner with an inclination of 20^o and

it is reflected with the same angle. Moreover, the reflection coefficient is 1.

In order to check the efficiency of the method, the relative error of the amplitude is calculated at a particular node. Firstly, the relative error is calculated for the amplitude of the incident wave without considering the reflection and transmission problem and next, the relative error for the amplitude of the reflected wave is obtained. The values of these relative errors are 0.6% and 1.1%, respectively, at the node located on $P = (1.05, -0.375)$.

Figure (1) shows the displacement along the Y axis for all nodes in the domain at a fixed time, $t = 0.75$ s, and figure (2) displays the displacement along the Y axis at a fixed node, $P = (1.05, -0.375)$, for the temporal interval.

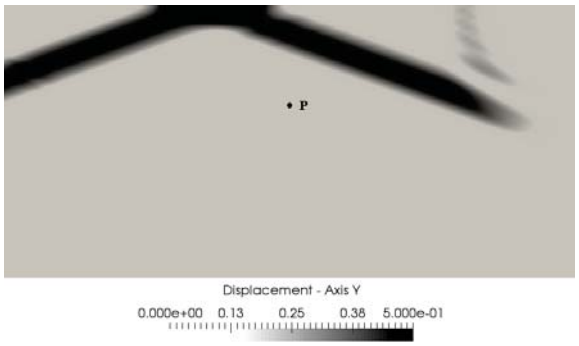


Figure 1: Incident and reflected SH wave at time $t = 0.75$ s

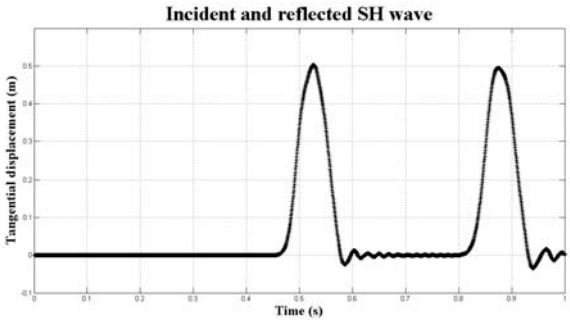


Figure 2: Incident and reflected SH wave at node $P = (1.05, -0.375)$. The right lobe is due to the incident wave and the left one is due to the reflected wave

4.2 Welded interface

The domain $D = [0, 2.1] \times [-1.5, 0] \subset \mathbb{R}^2$ is considered and a discretization of 56 481 nodes is performed. The lower medium, $D_1 = [0, 2.1] \times (-0.75, 0]$, has $\lambda_1 = 1$ Pa, $\mu_1 = 1$ Pa and $\rho = 1\ kg/m^3$, so $\beta_1 = 1\ m/s$, and the upper medium, $D_1 = [0, 2.1] \times [-1.5, -0.75)$, has $\lambda_2 = 1$ Pa, $\mu_2 = 4$ Pa and $\rho_2 = 1\ kg/m^3$, so $\beta_2 = 2\ m/s$. The incident SH plane wave arrives as described for the previous case and so, the angles of reflection and transmission are 20^o and 43.16^o, respectively, and the reflection and transmission coefficients are 0.4408 and 1.4408, respectively, so the amplitude of the reflected wave is $A_1 = 0.2204$ m and the amplitude of the transmitted wave is $A_2 = 0.7204$ m.

The values of the relative errors are 1.8%, at the node located on $P_1 = (1, -0.5)$, and 2.2%, at the node located on $P_2 = (1, -1)$, for the transmitted and reflected wave, respectively.

Analogously as for the previous case, figure (3) displays the displacement at time $t = 1.25$ s and figure (4) displays the displacements at nodes P_1 and P_2, respectively.

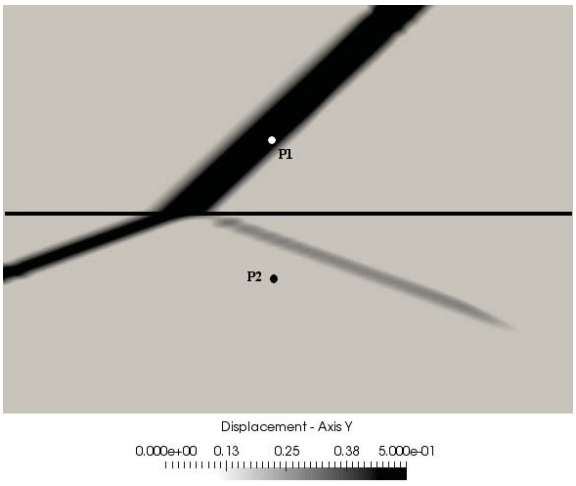


Figure 3: Incident, reflected and transmitted SH wave at time $t = 1.25$ s

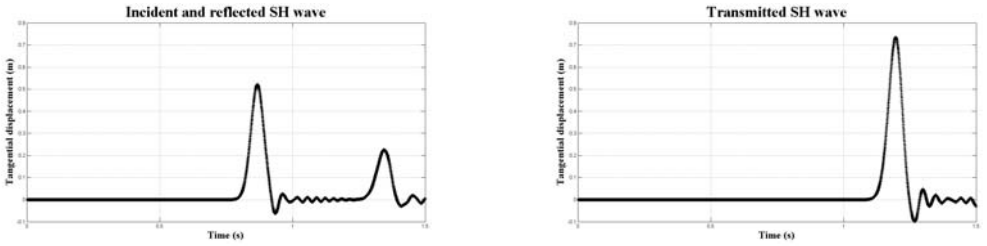


Figure 4: Left: Incident and reflected SH wave at node $P_1 = (1, -0.5)$. The right lobe is due to the incident wave and the left one is due to the reflected wave. Right: Transmitted SH wave at node $P_2 = (1, -1)$.

5 Conclusions

A matrix formulation of the generalized finite difference method is introduced and a necessary and sufficient condition for the uniqueness of the solution is proved. The GFD scheme for SH wave is obtained and its stability is analysed. Heterogeneous approximation is chosen and no nodes are placed at the interface of two homogeneous media. Two academic examples in heterogeneous media are developed: free surface case and welded interface pretend the Earth's surface and Earth's interior, respectively. Results show, the efficiency of the method and its accuracy.

Acknowledgements

The authors acknowledge the support of the Escuela Técnica Superior de Ingenieros Industriales (UNED) of Spain, project Ref: 2015-IFC02.

References

[1] T Belytschko, Y Krongauz, D Organ, M Fleming, P Krysl. Meshless methods: An overview and recent developments, Comput.Methods Appl.Mech.Eng. 139 (1996) 3-47.

[2] J Orkisz. Finite difference method (Part III), Handbook of Computational Solid Mechanics. (1998) 336-432.

[3] JJ Benito, F Ureña, L Gavete. Influence of several factors in the generalized finite difference method, Appl.Math.Model. 25 (2001) 1039.

[4] F Ureña, JJ Benito, E Salete, L Gavete. A note on the application of the generalized finite difference method to seismic wave propagation in 2D, J.Comput.Appl.Math. 236 (2012) 3016.

[5] JJ Benito, F Ureña, L Gavete, E Salete, A Muelas. A GFDM with PML for seismic wave equations in heterogeneous media, J.Comput.Appl.Math. 252 (2013) 40.

[6] JJ Benito, F Ureña, E Salete, A Muelas, L Gavete, R Galindo. Wave propagation in soils problems using the Generalized Finite Difference Method, Soil Dyn.Earthquake Eng. 79, Part A (2015) 190-198.

[7] S Stein, M Wysession, An introduction to seismology, earthquakes, and earth structure, John Wiley & Sons 2009.

[8] A Ilan, A Ungar, Z Alterman. An improved representation of boundary conditions in finite difference schemes for seismological problems, Geophysical Journal International. 43 (1975) 727-745.

[9] P Moczo, J Kristek, V Vavryuk, RJ Archuleta, L Halada. 3D heterogeneous staggered-grid finite-difference modeling of seismic motion with volume harmonic and arithmetic averaging of elastic moduli and densities, Bulletin of the Seismological Society of America. 92 (2002) 3042-3066.

[10] MC Bartholomew-Biggs, Nonlinear optimization with engineering applications, Springer, New York, 2008.

[11] AR Mitchell, DF Griffiths, The finite difference method in partial differential equations, John Wiley 1980.

[12] A Ben-Menahem, SJ Singh, Seismic waves and sources, Springer Science & Business Media 2012.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Optimizing a pivot-based algorithm for similarity search on a GPU-based platform

Roberto Uribe-Paredes[1], Enrique Arias[2], Diego Cazorla[2] and José L. Sánchez[2]

[1] *Computer Engeneering Dept., University of Magallanes*
Ave. Bulnes 01855, Punta Arenas, Chile.
Tel.: +56-61-207007; Fax: +56-61-232284

[2] *Computing Systems Dept., University of Castilla-La Mancha*
Campus Universitario s/n, Albacete, Spain.
Tel.: +34-967-599200; Fax: +34-967-599224

emails: roberto.uribeparedes@gmail.com, enrique.arias@uclm.es, diego.cazorla@uclm.es, jose.sgarcia@uclm.es

Abstract

In order to obtain an efficient implementation of a given algorithm for any hardware platform, it is very important to take into account the features of that platform when designing such as implementation. This is especially true in the case of GPU-based platforms. In this paper we show how different properties of current GPUs are exploited for improving a version of the General Metric Structure similarity search algorithm introduced by the authors. The optimizations are introduced step by step in a reasoned way and are analysed from performance point of view.

Key words: Similarity search, metric spaces, GPU-based platforms, optimization

1 Introduction

The search of similar objects in a large collection of objects stored in a metric database has become a very interesting problem in the last decades, due to the fact that this kind of search can be found in different applications such as voice and image recognition, data mining, and many others. Similarity is modeled in many interesting cases through metric spaces, and the search of similar objects through range search or nearest neighbors (kNN).

A metric space $(\mathbb{X}, d)$ is a set $\mathbb{X}$ and a distance function $d : \mathbb{X}^2 \to \mathbb{R}$, so that $\forall x, y, z \in \mathbb{X}$ fulfills the properties of positiveness $[d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y]$, symmetry $[d(x, y) = d(y, x)]$ and triangle inequality $[d(x, y) + d(y, z) \geq (d(x, z)]$.

There exist diverse metric data structures and indices for reducing the number of distance evaluations in the search process. Basically, metric data structures can be grouped into two classes: *clustering*-based method (*BST*, *GHT*, *M-Tree*, *EGNAT*, and many others) and *pivots*-based method (*LAESA* [1], *FQT* and its variants [2], *Spaghettis* and its variants [3], *FQA* [4], *SSS-Index* [5] and others), in which this work is focused on. Most of these solutions for similarity search are described and analyzed in [6, 7].

However, these metric data structures are very evolutioned and optimized for sequential platforms where the code is executed on the CPU using, mainly, main memory (some of them are optimized for secondary memory). On the other hand, in order to accelerate the search process, GPUs are being used, but most proposals are focused on force brute or kNN approaches [8, 9], and few solutions can be found on range search methods [10, 11].

The authors use a generic version of a pivot-based structure called Generic Metric Structure (GMS), and have previously obtained an implementation of this method for GPU-based platforms, but using old GPUs and applying few optimizations. In this work, we have used a more advanced graphic card and applied a larger set of optimizations. This constitutes the main contribution of this work.

The paper is structured as follows: Section 2 briefly introduces the Generic Metric Data structure and the sequential implementation for solving the similarity search. In Section 4 the experimental environment is described, and Section 3 presents the different approaches towards the optimized GPU implementation. Finally, Section 5 outlines the conclusions of this work and the future work to address.

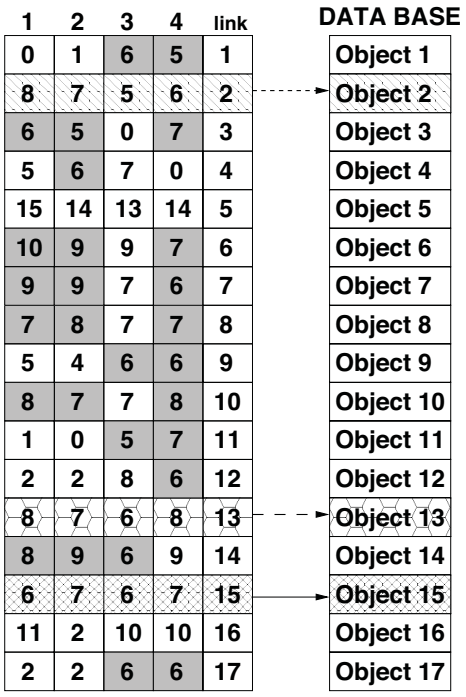
2 A Generic Metric Structure (GMS)

GMS structure is implicitly present in the literature and the authors proved its suitability for implementation on GPU-based plataforms [11]. This generic metric structure (GMS) consists of a bidimensional array data structure (Fig. 1b) whose dimensions are $N \times P$, being N the number of objects and P the number of pivots. In general terms, other popular metric data structures as Spaghettis [3] and SSS-Index [5] could be considered as generic bidimensional array data structures. The difference between these structures is the way of obtaining the pivots or the way in which the structure is stored. GMS avoids the sorting process (against the Spaghettis method) because this process is computationally expensive on GPUs, and it also avoids the selection of the pivots which is carried out randomly (against the SSS-Index selection).

For GMS, the searching process, given a query q and a range r, is carried out according to the following steps:

```
1: for i = 1 to num_queries do
2:    for k = 1 to num_pivots do
3:       d_k ← distance(q_i,p_k)
4:    end for
5:    for j = 1 to num_objects do
6:       discarded ← false
7:       for k = 1 to num_pivots do
8:          if d_k − r > s_jk || d_k + r < s_jk then
9:             discarded ← true
10:            break;
11:         end if
12:      end for
13:      if !discarded then
14:         if distance(x_j,q_i) ≤ r then
15:            add to result
16:         end if
17:      end if
18:   end for
19: end for
```

(a) Sequential GMS algorithm



(b) GMS structure

Figure 1: Sequential GMS algorithm and example of a GMS structure

1. From the distance between q and all pivots $p_1, \ldots, p_k$ we obtain k intervals in the form $[a_1, b_1], ..., [a_k, b_k]$, where $a_i = d(p_i, q)$ - r and $b_i = d(p_i, q)$ + r.

2. The objects in the intersection of all intervals are candidates to the query q.

3. For each candidate object y, the distance $d(q, y)$ is calculated, and if $d(q, y) \leq r$, then the object y is a solution to the query q.

Fig. 1a shows the pseudocode corresponding to the searching process using GMS. Fig. 1b shows how GMS structure is built using 4 pivots to index a database of 17 objects [12]. For a query q with distances to pivots $d(q, pi) = 8, 7, 4, 6$ and a search range $r = 2$, define the intervals $(6, 10), (5, 9), (2, 6), (4, 8)$ over which the searching is going to be carried out. The objects $2, 13, 15$ are candidates and their real distance to the query must be calculated.

Note that this structure is previously created just once by applying the distance function considering all the objects of the database and for all the pivots.

3 GPU-based implementation of the GMS algorithm

As we can observe from pseudocode in Fig. 1a, each query is solved independently of the others, and therefore the sequential algorithm is fully parallelizable. At first, we begin with a simple parallel implementation, where all data is stored in global memory. The main actions of this version (Fig. 2) are device memory allocation, information transferences between host and device memory, and definition and invocation of the kernel.

```
1: . . .
2: cudaMalloc(dbGPU,dSize);
3: cudaMalloc(queryGPU,qSize);
4: cudaMalloc(pivotsGPU,pSize);
5: cudaMalloc(gmsGPU,gSize);
6: cudaMalloc(resultGPU,rSize);
7: . . .
8: cudaMemcpy(dbGPU,dbCPU,dSize,cudaMemcpyHostToDevice);
9: cudaMemcpy(queryGPU,queryCPU,qSize,cudaMemcpyHostToDevice);
10: cudaMemcpy(pivotsGPU,pivotsCPU,pSize,cudaMemcpyHostToDevice);
11: cudaMemcpy(gmsGPU,gmsCPU,gSize,cudaMemcpyHostToDevice);
12: . . .
13: dim3 grid( );
14: dim3 threads( );
15: kernel<<<grid,threads>>>(dbGPU,queryGPU,pivotsGPU,gmsGPU,resultGPU);
16: cudaThreadSynchronize( );
17: . . .
18: cudaMemcpy(resultCPU,resultGPU,rSize,cudaMemcpyDeviceToHost);
19: . . .
```

Figure 2: General scheme of the GPU-based implementation for the GMS algorithm

3.1 Task definition and grid configuration

In order to complete the kernel, the task performed by each thread has to be defined. In this sense, several alternatives can be considered, each one leading to a different granularity level. We have studied the following three cases:

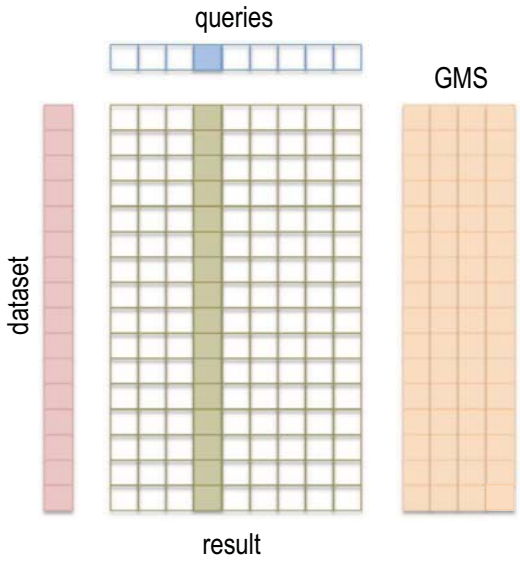
A) Each thread completely solves a query, that is, each thread determines which objects of the database are a solution for a given query (Fig. 3). The invocation of this kernel generates as many threads as queries.

 ISBN: 978-84-608-6082-2

```
__global__ void kernel-A (db,query,piv,gms,result)
int tx = threadIdx.x;
int bx = blockIdx.x;
int index = bx*blockDim.x+tx;
int pos;
for pos=0 to num_objects − 1 do
  lines 6 to 13 of code in Fig. 1
  if distance(db[pos],query[index])≤r then
    result[pos,index] = pos;
  end if
end for

dim3 grid(num_queries/BlockSize)
dim3 threads(BlockSize)
```

(a) Kernel and grid definition



(b) Final results

Figure 3: Case A: kernel and data managed by a given thread

B) As many threads as objects in the database are created, and each thread detemines if a given object is solution for any query (Fig. 4).

C) Several threads collaborate to solve a query, that is, each thread determines if a given object of the database is a solution to a given query (Fig. 5). The number of threads executing the kernel is given by the product $num\_objects \times num\_queries$.

In Figs. 3, 4, and 5, Subfigure (a) includes the kernel and grid definition, and Subfigure (b) indicates the data managed by the threads. For instance, in case A, each thread reads a query and all objects in the database, and writes in *result* structure the result of checking out if those objects are solution for the query. Case B is similar, and in case C, each thread reads a query and an object, and writes in *result* structure if the object is solution for the query.

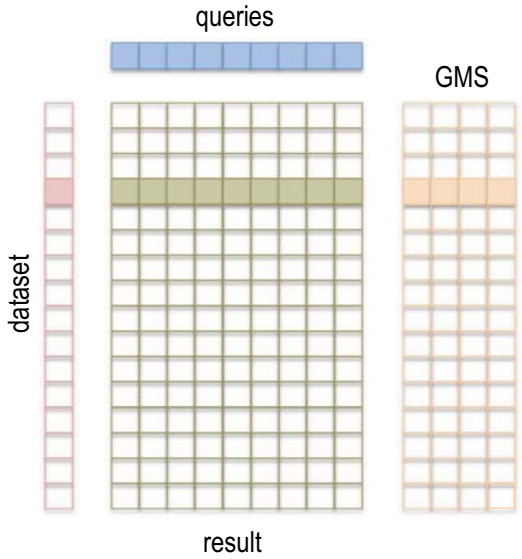
Since our purpose is to compare different versions of the same algorithm under the same conditions, two assumptions have been considered for the three implementations: The database can be completely allocated in the device memory, and only one kernel is considered and all the queries are solved in just one call to this kernel.

On the other hand, fixed the number of threads in each case, and considering one dimension both for grid and thread blocks definition, several thread-block sizes can be

```
_global_ void kernel-B (db,query,piv,gms,result)
int tx = threadIdx.x;
int bx = blockIdx.x;
int index = bx*blockDim.x+tx;
int pos;
for pos=0 to num_queries − 1 do
  lines 6 to 13 of code in Fig. 1
  if distancie(db[index],query[pos])≤r then
    result[index,pos] = index;
  end if
end for

dim3 grid(num_objects/BlockSize)
dim3 threads(BlockSize)
```

(a) Kernel and grid definition



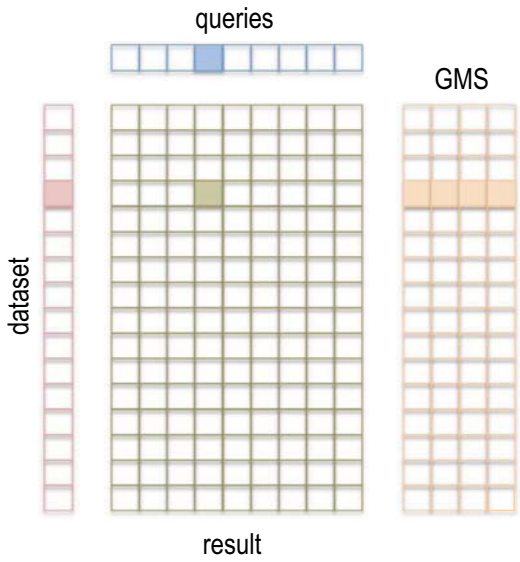
(b) Final results

Figure 4: Case B: kernel and data managed by a given thread

```
_global_ void kernel-C (db,query,piv,gms,result)
int tx = threadIdx.x;
int ty = threadIdx.y;
int bx = blockIdx.x;
int by = blockIdx.y;
int indexX = bx*blockDim.x+tx;
int indexY = by*blockDim.y+ty;
lines 6 to 13 of code in Fig. 1
if distance(db[indexY],query[indexX])≤r then
  result[indexY,indexX] = indexY;
end if

dim3 grid(num_queries/BlockSize,
          num_objects/BlockSize)
dim3 threads(BlockSize, BlockSize)
```

(a) Kernel and grid definition



(b) Final results

Figure 5: Case C: kernel and data managed by a given thread

considered. We have varied this size, through the parameter *BlockSize* in the codes, from 32 to 1024 in order to find the most appropriate value for each case.

4 Experimental results

We have used a CPU-GPU platform consisting of the following main components: The host is compound by an Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz, 256 GB of DDR3 memory and 2 MB of chache memory. On the other hand, the device is composed of an NVIDIA GPU Tesla K40m based on Kepler architecture. It has 2,880 cores divided in 15 multiprocessors with 192 cores each, and also 49Kb shared memory and 12GB of device memory. Host and device are connected by a PCI-E 3.0 bus. Regarding programming environment, we used NVIDIA CUDA 7.5 toolkit.

We have considered two datasets: a subset of the Spanish dictionary using the edit distance and, for each query, a range search between 1 and 4 was considered. Given two words, this edit distance is defined as the minimum number of insertions, deletions or substitutions of characters needed to make one of the words equal to the other. The second space is the CoPhIR database (Content-based Photo Image Retrieval) [13] which includes images from metadata from Flickr with five MPEG-7 descriptors. We have used the 64 dimensional vector Color Structure descriptor for each image, which is an usual dimension size [14]. Any quadratic form can be used as a distance, thus we chose Euclidean distance as the simplest meaningful alternative. The radius used was that allowing to retrieve 0.01%, 0.1% and 1% from the dataset.

In the case of the dictionary, two sizes of the database have been considered (65536 and 16384 objects), and three sets of queries (1024, 2048 and 4096). In the case of CoPhIR database, we have used three sizes of the database (262144, 131072 and 65536 objects), and two sets of queries (4096 and 8192). We select 32 pivots to build the GMS because this amount of pivots gives the best results [12, 11]. Pivots were randomly selected. The time spent during the construction of the structure is not computed, this is considered as a preprocessing time.

For all the experiments, we only consider the time of the searching process, including CPU-GPU data transfer time. Each value shown in the figures is the average of 10 different executions of the algorithm. The computation of the different structures, for instance the pivots, is computed offline. In this way, all the methods are in the same conditions in order to carry out a correct comparison.

4.1 Task granularity

Fig. 6 shows the kernel time for some of the possible configurations we can obtain considering all varied parameters. In this case, as the baseline code (Fig. 2) is the same for the three cases and the differences between them are in the kernel and the number of threads,

for comparative purposes we only consider the performance results obtained by the kernel. We have varied the database size, the number of queries, range and percentage of retrieved objects in the dataset, and we have obtained the same tendency in the results. In the case of Spanish dictionary, for all the tests carried out, the block size offering the best results is 256 threads per block in most cases, which corresponds to 100% occupancy in the GPU. In some parameter configurations the same results and GPU occupancy are also obtained for 128 and 512 threads per block. However, for CoPhIR dataset, 1024 threads per block offers the best results.

On the other hand, clearly, the worst results are obtained by the kernel in case A, whereas kernels B and C have a similar behavior, for the best thread block sizes. These kernels generate a high number of threads, and as a consequence the underlying hardware is better exploited. According to these results, from now we only consider kernel C, and 256 and 1024 threads per block (dictionary and CoPhIR, respectively) for the remaining experiments.

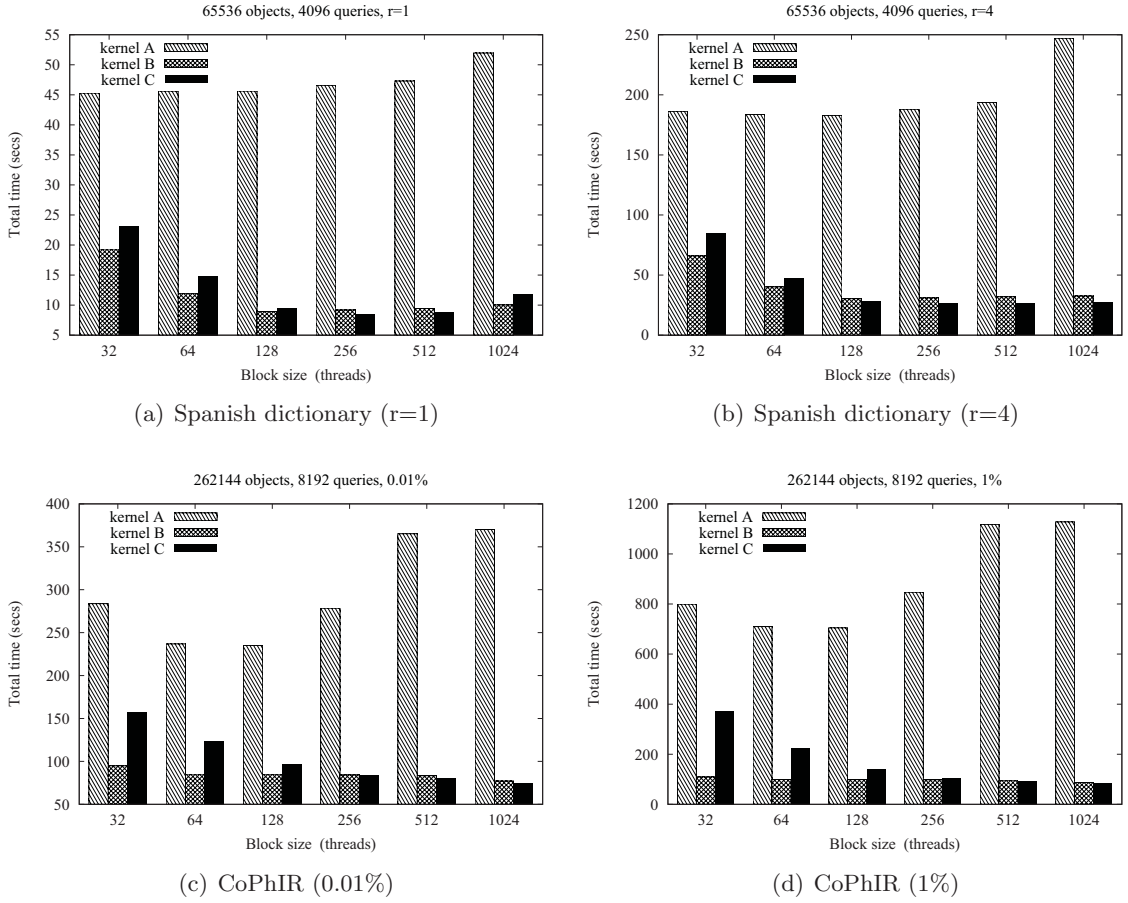
Regarding throughput, and if we consider the CoPhIR dataset for the best results in Fig. 6a, we obtain 6.26 and 6.95 Gflop/s for kernels B and C, respectively. These results are far from the theoretical ones that the GPU can provide. To reason about possible causes let's look at the *ptx* code of the kernel C. In Fig. 7 we have included the key instructions, which correspond to the distance function. As mentioned in this section, this function is the Euclidean distance and the code in Fig. 7 basically implements the equation $\sqrt{\sum_{i=1}^{112}(x_i - y_i)^2}$, where x, y are vectors.

From the *ptx* code it is possible to estimate the performance of the kernel in terms of floating point operations per time unit. We can observe that this kind of operations are `sub`, `fma` (sum and multiplication) and `sqrt`, but in order to calculate the performance it is only necessary to consider the loop B1 because the result is virtually the same that if all instructions are considered. In that loop there are 10 operations (8 single-operation instructions and 2 operations for `fma` instruction) and only three of them are useful. Therefore, the kernel C in the best case could achieve up to 648 Gflop/s (2880 cores $\times$ 750 MHz $\times$ 3/10 flop), far from performance values that actually have been obtained.

Note that in the loop there are also two load operations. We can calculate the pressure of the cores over the device memory to complete the required memory accesses to carry out all the distance evaluations. We obtain 1728 GB/s (2880 cores $\times$ 750 MHz $\times$ 2/10 loads $\times$4 bytes/load), that is, to complete all distance evaluations a bandwidth of 1728 GB/s is required, meanwhile the available bandwidth is 288 GB/s. As a consequence the device memory represents a bottleneck to obtain the peak performance of the kernel code.

4.2 Memory latency reduction

In order to reduce the overhead produced by the accesses to global memory of the device, keeping the basic kernel scheme, several strategies can be adopted, as for example to exploit



(a) Spanish dictionary (r=1)

(b) Spanish dictionary (r=4)

(c) CoPhIR (0.01%)

(d) CoPhIR (1%)

Figure 6: Time for kernels A, B and C when varying threads block size

cache memory or to use shared memory.

A first modification consists in storing the GMS structure in a different way, that is transposed, to reduce the time when it is read. The improvement varies between 5% and 5% for CoPhIR and Spanish dictionary datasets, respectively. If we use shared memory to store a group of queries, objects or distances in GMS, depending on the kernel, more significant reductions on the total time are achieved: Up to 35% in the CoPhIR case, and 22% for Spanish dictionary case (label `Shared` is used in Fig. 8). These gains are respect to global-memory-GPU GMS algorithm.

Fig. 8 shows speedup values with respect to the time for the GMS sequential algorithm. It must be also taken into account that each improvement in Fig. 8 includes the previous

```
   add.s64       %rd17,%rd12,%rd13;     cvta.to.global.u64  %rd14,%rd7;
   mov.u32       %r24,0;                cvt.f64.f32         %fd1,%f9;
   mov.f32       %f9,0f00000000;        sqrt.rn.f64         %fd2,%fd1;
                                        cvt.rn.f32.f64      %f8,%fd2;
B1:                                     setp.gtu.f32        %p2,%f8,%f3;
  ld.global.f32 %f5,[%rd18];            add.s32             %r19,%r1,%r4;
  ld.global.f32 %f6,[%rd17];            selp.b32            %r20,-1,%r19,%p2;
  sub.f32       %f7,%f6,%f5;            add.s32             %r21,%r2,%r3;
  fma.rn.f32    %f9,%f7,%f7,%f9;        mad.lo.s32          %r22,%r19,%r7,%r21;
  add.s64       %rd18,%rd18,4;          mul.wide.s32        %rd15,%r22,4;
  add.s64       %rd17,%rd17,4;          add.s64             %rd16,%rd14,%rd15;
  add.s32       %r24,%r24,1;            st.global.u32       [%rd16],%r20;
  setp.ne.s32   %p1,%r24,112;
  @%p1 bra      B1;
```

Figure 7: Ptx code of the kernel C

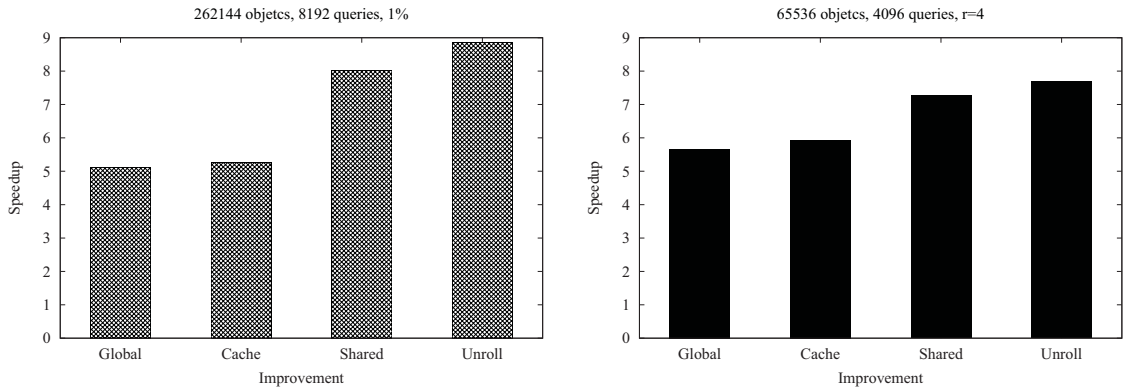
ones.

4.3 Executed instruction reduction

Another way to reduce the kernel time consists in eliminating those operations which are not part of the core data computation, such as branches and address calculations. In order to do this, we can apply loop unrolling, a classical compiler optimization. Since the number of times that the loops are unrolled affects the number of registers used per thread and, as a consequence, the number of thread blocks that can be scheduled per multiprocessor, we have conducted several experiments to determine the best number for each kernel. The results have been also included in Fig. 8 using label `Unroll`.

5 Conclusions and the future work

This work shows how to obtain successive improvements on a particular algorithm applying different techniques based on the knowledge of the underlying architecture. We have analysed the effect on the performance of task granularity, threads block configuration, cache memory, shared memory or loop unrolling, and we have obtained improvements when all together have been considered. Specifically, reductions on time of nearly 9x have been achieved for CoPhIR and Spanish dictionary cases, with respect to the best sequential algorithm. We would like to remark that the different alternatives presented in this paper can be applied to other pivots-based structures, especially, those based on distance tables, as LAESA, SSSIndex, and others. Therefore, we intend to utilize the kernels developed



(a) CoPhIR (b) Spanish dictionary

Figure 8: Results of speedup after applying different improvements to kernel C, with respect to GMS sequential algorithms.

for those structures. Our intention is also to improve the kernels taking advantage of the improvements provided by the architectures of new GPUs, mainly in the line of reducing memory latency.

Acknowledgements

This work has been partially supported by the MINECO under the project CGL2013-48367-P, and by the JCCM and European Commission (FEDER funds) under the project PEII-2014-028-P.

References

[1] Micó, M.L., Oncina, J., Vidal, E.: A new version of the nearest-neighbour approximating and eliminating search algorithm (aesa) with linear preprocessing time and memory requirements. Pattern Recognition Letters **15**(1) (January 1994) 9–17

[2] Baeza-Yates, R., Cunto, W., Manber, U., Wu, S.: Proximity matching using fixed-queries trees. In: 5th Combinatorial Pattern Matching (CPM'94). Volume 807 of LNCS. Springer Berlin Heidelberg (1994) 198–212

[3] Chávez, E., Marroquín, J.L., Baeza-Yates, R.: Spaghettis: An array based algorithm for similarity queries in metric spaces. In: 6th International Symposium on String Processing and Information Retrieval (SPIRE'99), IEEE CS Press (1999) 38–46

[4] Chávez, E., Marroquín, J.L., Navarro, G.: Fixed queries array: A fast and economical data structure for proximity searching. Multimedia Tools and Applications **14**(2) (June 2001) 113–135

[5] Pedreira, O., Brisaboa, N.R.: Spatial selection of sparse pivots for similarity search in metric spaces. In: 33rd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2007). Volume 4362 of LNCS., Harrachov, Czech Republic, Springer (2007) 434–445

[6] Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.L.: Searching in metric spaces. ACM Computing Surveys **33**(3) (2001) 273–321

[7] Hetland, M.: The basic principles of metric indexing. In Coello, C., Dehuri, S., Ghosh, S., eds.: Swarm Intelligence for Multi-objective Problems in Data Mining. Volume 242 of Studies in Computational Intelligence. Springer Berlin / Heidelberg (2009) 199–232

[8] Kuang, Q., Zhao, L.: A practical GPU based kNN algorithm. International Symposium on Computer Science and Computational Technology (ISCSCT) (2009) 151–155

[9] Garcia, V., Debreuve, E., Barlaud, M.: Fast k nearest neighbor search using GPU. Computer Vision and Pattern Recognition Workshop **0** (2008) 1–6

[10] Barrientos, R.J., Gómez, J.I., Tenllado, C., Matias, M.P., Marin, M.: Range query processing on single and multi GPU environments. Computers & Electrical Engineering **39**(8) (2013) 2656 – 2668

[11] Uribe-Paredes, R., Arias, E., Sánchez, J.L., Cazorla, D., Valero-Lara, P.: Improving the performance for the range search on metric spaces using a multi-GPU platform. In: Database and Expert Systems Applications (DEXA). Volume 7447 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 442–449

[12] Uribe-Paredes, R., Valero-Lara, P., Arias, E., Sanchez, J.L., Cazorla, D.: Similarity search implementations for multi-core and many-core processors. In: International Conference on High Performance Computing and Simulation (HPCS). (2011) 656–663

[13] Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Piccioli, T., Rabitti, F.: CoPhIR: a test collection for content-based image retrieval. CoRR **abs/0905.4627v2** (2009)

[14] Kruliš, M., Skopal, T., Lokoč, J., Beecks, C.: Combining cpu and gpu architectures for fast similarity search. Distributed and Parallel Databases **30**(3-4) (2012) 179–207

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Contour curves and isophotes on ruled surfaces

Jan Vršek[1,2]

[1] *NTIS – New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

[2] *Department of mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

emails: vrsekjan@kma.zcu.cz

Abstract

The ruled surfaces, i.e., surfaces generated by one parametric set of lines, are widely used in the field of applied geometry. Isophote on a surface is a curve consisting of surface points whose normals form a constant angle with some fixed vector. Choosing an angle equal to $\pi/2$ we obtain a special instance of isophote – the so called contour curve. While contours on rational ruled surfaces are rational curves, this is no longer true for the isophotes. Hence we will provide a formula for their genus. Moreover we will show that the only surfaces with a rational generic contour are just rational ruled surfaces and a one particular class of cubic surfaces. In addition we will deal with the reconstruction of ruled surfaces from their contours and silhouettes.

Key words: Contour curve, isophote, ruled surface, surface reconstruction

1 Introduction

The aim of the paper is the study of contour curves and isophotes on rational ruled surfaces. These characteristic curves of the surface are usually studied via methods of differential geometry [5, 6, 9]. Since people in the geometric modelling community prefer to work with rationally parametrizable objects, the description of rational contours and isophotes on surfaces is needed. See e.g. [1] for isophotes on surfaces with rational support function or [3] for the usage of rational contours in parameterization problem. We will show that rational ruled surfaces are almost the only surfaces whose generic contour curve is rational. Given a contour curve or silhouette of the surface there is a natural question whether we

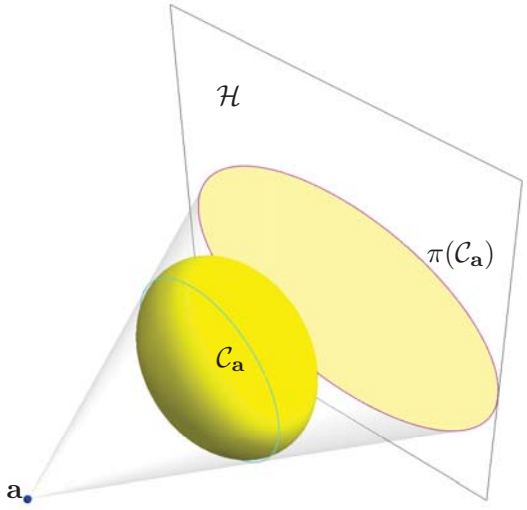


Figure 1: The contour $\mathcal{C}_\mathbf{a}$ (cyan) w.r.t. point $\mathbf{a}$ on the surface. And the silhouette $\pi(\mathcal{C}_\mathbf{a})$ (magenta) as the projection of the contour into the plane $\mathcal{H}$.

are able to reconstruct the surface from it. We will answer it in full for the class of rational ruled surfaces. Unlike the contours the isophotes on rational ruled surfaces are not rational any more. More precisely they are always hyperelliptic and we will find a formula for its genus. Because of the used methods we decided to work in projective space over the field of complex numbers.

Let $\mathcal{X}$ be a surface in the projective space $\mathbb{P}^3_\mathbb{C}$ and $\mathcal{X}_{sm}$ denote the set of its smooth points. Then for any fixed point $\mathbf{a}$ the *contour* $\mathcal{C}_\mathbf{a}$ of $\mathcal{X}$ with respect to a viewpoint $\mathbf{a}$ is defined as the closure of the set

$$\{\mathbf{p} \in \mathcal{X}_{sm} : \ \mathbf{a} \in T_\mathbf{p}\mathcal{X}_{sm}\}, \tag{1}$$

where $T_\mathbf{p}\mathcal{X}_{sm}$ denotes the tangent plane at $\mathbf{p}$. If $\mathcal{H}$ is an arbitrary plane not passing through $\mathbf{a}$ then we may project a contour $\mathcal{C}_\mathbf{a}$ from the point $\mathbf{a}$ to the plane $\mathcal{H}$ to obtain the so called *silhouette*, see Fig. 1.

Let $(x_0 : x_1 : x_2 : x_3)$ be coordinates in $\mathbb{P}^3_\mathbb{R}$. Fix a hyperplane $\omega : x_0 = 0$ and the absolute conic section $\Omega : x_0 = x_1^2 + x_2^2 + x_3^2$ then the complement $\mathbb{A}^3_\mathbb{R} = \mathbb{P}^3_\mathbb{R} \backslash \omega$ is an affine space endowed with usual Euclidean metric. The plane ω is called a plane at infinity and its points can be understood as directions in $\mathbb{A}^3_\mathbb{R}$. Depending on the position of the point $\mathbf{a}$ we distinguish between contour w.r.t central projection ($\mathbf{a} \notin \omega$) and parallel projection ($\mathbf{a} \in \omega$). Unlike the contour the definition of isophote depends on the metric of ambient space. In differential geometry, the isophote is defined as a loci of points where the surface normals

encloses a constant angle with fixed vector. This definition is not suitable when attacking the problem with algebraic techniques for two reasons. First, we would like to define them over $\mathbb{C}$ and not $\mathbb{R}$. And second, with this definition the isophote is not an algebraic curve but only its half. Hence we define the *isophote* $\mathcal{I}_{\mathbf{a},\alpha}$ to be the closure of the set

$$\left\{\mathbf{p} \in \mathcal{X}_{sm} : (\mathbf{n}(\mathbf{p}) \cdot \mathbf{a})^2 = \alpha^2 \mathbf{n}(\mathbf{p}) \cdot \mathbf{n}(\mathbf{p})\right\}, \tag{2}$$

where $\mathbf{a} = (a_1, a_2, a_3)$, $\sum a_i^2 = 1$, is a given direction vector and $\phi = \arccos \alpha$ is the angle. Hence $\mathcal{I}_{\mathbf{a},\alpha}$ is a set of points on the surfaces where the normal line forms an angle $\pm\phi$ with direction $\mathbf{a}$. Considering $\mathbf{a}$ to be a point in ω, i.e. $\mathbf{a} = (0 : a_1 : a_2 : a_3)$ and letting $\alpha = \cos \pi/2 = 0$ we obtain exactly a contour curve $\mathcal{C}_{\mathbf{a}} = \mathcal{I}_{\mathbf{a},0}$.

One of the simplest class of rational parametric surfaces used in geometric modelling are the quadratic patches. These are the projections of Veronese surface to $\mathbb{P}^3_{\mathbb{C}}$. Depending on the projection (or equivalently on the number of base points) the quadratic patch parameterizes one of the following surfaces.

1. quadric,
2. ruled cubic with double line,
3. Steiner surface (of degree 4).

As we will see, the regular quadrics have rational contours. The same is true for ruled cubic as contour curves on every rational ruled surface are rational. However a generic projection of Veronese and thus almost all quadratically parameterized surfaces in $\mathbb{P}^3_{\mathbb{C}}$ are Steiner quartics. Their generic contour curves are elliptic curves and thus they are not rational. Hence even a very simple surface does not posses a contours parameterizable by the standard techniques used in CAGD. Nevertheless it is known that contours on rational ruled surfaces are rational, see e.g. [7]. The following theorem completes the lists of all non-developable surfaces with this property..

Theorem 1.1 *A generic contour curve on a surface in $\mathbb{P}^3_{\mathbb{C}}$ is rational if and only if the surface is rational ruled or the Cayley cubic, i.e. rational cubic surface with four double points.*

Let $\mathcal{C} \subset \mathbb{P}^3_{\mathbb{C}}$ be a curve defined as simultaneous solution to homogeneous equations $G_i(x_0, x_1, x_2, x_3) = 0$ for $i = 1, \ldots, m$ and let $\mathbf{a} \in \mathbb{P}^3_{\mathbb{C}}$ be a fixed point. A surface $\mathcal{X}$ (not necessarily rational and ruled at this moment) containing the curve $\mathcal{C}$ as a contour curve w.r.t. a point $\mathbf{a}$ possesses a defining polynomial $F = \sum_{i=1}^m H_i G_i$ for some homogeneous polynomials H_i. It can be proved that these polynomials must fulfil the relation

$$\sum_{i=1}^{m} H_i \left(\sum_{j=0}^{3} \frac{\partial G_i}{\partial x_j} a_j \right) - \sum_{i=1}^{m} L_i G_i \equiv 0, \tag{3}$$

 ISBN: 978-84-608-6082-2

for some homogeneous L_i for $i = 1, \ldots, m$. This leads to the study of the so called syzygy modules, see e.g. [4]. It turns out that a surface in $\mathbb{P}^3_{\mathbb{C}}$ is uniquely determined by finitely many contour curves.

Problem 1.2 *For a given centers of projection* $\mathbf{a}_i$ *and curves* $\mathcal{C}_i$ *for* $i = 1, \ldots, k$ *find the unique surface with these curves as contours, i.e.* $\mathcal{C}_i = \mathcal{C}_{\mathbf{a}_i}$.

The above demonstration indicates that the number k will be related to the degree of the sought surface. Whereas the exact value of k is not known in general, the knowledge about the geometry of the surface enables to reduce the number of needed contours, c.f. Theorem 3.2. Recall that the silhouette of a "2D image" of the surface is a projection of contour from the center $\mathbf{a}$ to some chosen plane. This justifies its name because it is just a silhouette of the "image" of the surface. Having enough such images it could be possible to reconstruct the original surface. Hence the following problem is motivated by a surface reconstruction.

Problem 1.3 *For a given centers of projection* $\mathbf{a}_i$ *and planar curves* $\mathcal{C}_i$ *for* $i = 1, \ldots, k$ *find the unique surface with these curves as silhouettes, i.e.* $\mathcal{C}_\alpha$ *is the projection of contour* $\mathcal{C}_{\mathbf{a}_i}$ *from the point* $\mathbf{a}_i$.

2 Quadrics

Although all quadrics are ruled surfaces we will treat them separately. There are two reasons for this. First, we can solve Problems 1.2 and 1.3 directly without a reference to the rulings. And second, the quadrics are a typical illustration of the drawback of an approach via complex numbers. Indeed the real part of sphere, paraboloid, etc. contains no line. So one cannot consider them to be ruled surfaces from the point of view of real geometry. Fortunately we may prove

Proposition 2.1 *Real rational ruled surface contains one–parametric set of real lines or it is a quadric.*

A regular quadric is defined be an equation $\mathbf{x}^T \cdot A \cdot \mathbf{x} = 0$ for some regular symmetric matrix A. The contour w.r.t. $\mathbf{a}$ is then the intersection of $\mathcal{Q}$ with its polar plane w.r.t. point $\mathbf{a}$, i.e. $\mathbf{a}^T \cdot A \cdot \mathbf{x} = 0$. It is a regular conic section whenever $\mathbf{a} \notin \mathcal{Q}$, or it consists of two lines otherwise. In order to solve Problem 1.2 it can be proved that for a fixed conic section $\mathcal{C}$ and a point $\mathbf{a}$ there exists a pencil of quadrics with this contour. The proposition then follows immediately.

Proposition 2.2 *A two contour curves uniquely determine a quadric.*

 ISBN: 978-84-608-6082-2

On the other hand given two compatible silhouettes there exists a one parametric set of solutions. Thus at least three of them are needed to reconstruct the quadric, c.f. Theorem 3.4

3 Ruled surfaces

3.1 Contours on ruled surfaces

Recall that by a *rational ruled surface* $\mathcal{R}$ is meant a rational surface in $\mathbb{P}^3_{\mathbb{C}}$ generated by a one-parametric set of lines – the so called rulings. Hence it admits a parameterization

$$\mathbf{x}(s,t) = \mathbf{p}(s) + t\mathbf{q}(s) = (p_0(s) + tq_0(s) : p_1(s) + tq_1(s) : p_2(s) + tq_2(s) : p_3(s) + tq_3(s)), \quad (4)$$

where $p_i(s)$ and $q_i(s)$ are polynomials. The rational curves $\mathbf{p}(s) = \mathbf{x}(s,0)$ and $\mathbf{q}(s) = \mathbf{x}(s,\infty)$ intersects a generic ruling exactly once. A curve on $\mathcal{R}$ with this property is called a section and it can be seen that each section is a rational curve. So a rational parametrization of the surface can be obtained by joining the corresponding points on two sections by line. If $\mathbf{p}(s)$ and $\mathbf{q}(s)$ does not intersect (there always exist such sections on $\mathcal{R}$) then the degree of the surface is $\deg \mathcal{R} = \deg \mathbf{p} + \deg \mathbf{q}$.

Denote $\mathbf{x} = (x_0 : x_1 : x_2 : x_3)$ and $\dot{\mathbf{p}}$, $\dot{\mathbf{q}}$ the derivatives of $\mathbf{p}$ and $\mathbf{q}$ respectively. Then the tangent plane at point $\mathbf{p}(s)$ is spanned by $\mathbf{p}(s)$, $\mathbf{q}(s)$ and $\dot{\mathbf{p}}(s)$, i.e. it possesses an equation $\det[\mathbf{x}, \mathbf{p}(s), \mathbf{q}(s), \dot{\mathbf{p}}(s)] = 0$ and analogously for $\mathbf{q}(s)$. Hence the condition on point $\mathbf{a}$ to be contained in the tangent plane of the surface can expressed

$$\det[\mathbf{a}, \mathbf{p}(s), \mathbf{q}(s), \dot{\mathbf{p}}(s) + t\dot{\mathbf{q}}(s)] = 0. \quad (5)$$

Expressing t from this equation and substituting back to (4) leads to the parameterization of the contour curve.

$$\mathbf{c_a}(s) = \det[\mathbf{a}, \mathbf{p}(s), \mathbf{q}(s), \dot{\mathbf{q}}(s)]\mathbf{p}(s) - \det[\mathbf{a}, \mathbf{p}(s), \mathbf{q}(s), \dot{\mathbf{p}}(s)]\mathbf{q}(s), \quad (6)$$

see [7] for the formula using Plücker coordinates. The contour curve is again a section, but not the one with minimal degree and thus two different contours on the ruled surface always intersect.

Lemma 3.1 *Two generic contours on rational ruled surface $\mathcal{R}$ intersect at $\deg \mathcal{R}$ regular non-torsal rulings and in the cuspidal points of regular torsal rulings.*

As well as in the quadratic case two contour curves are enough to determine the ruled surface. Let $\mathbf{c_a}(u)$ and $\mathbf{c_b}(v)$ be proper parameterizations of two contour lines $\mathcal{C}_\mathbf{a}$ and $\mathcal{C}_\mathbf{b}$ of the unknown surface $\mathcal{R}$. Thus there exist reparameterizations such that $\mathbf{c_a}(\phi(s))$ and $\mathbf{c_b}(\psi(s))$ lie on the same ruling for all s. To find this reparameterization, realise that a

tangent plane at point of $\mathcal{C}_{\mathbf{a}}$ is spanned by the tangent line to the contour and point $\mathbf{a}$. Moreover the tangent plane contains the ruling and thus the corresponding point on $\mathcal{C}_{\mathbf{b}}$ as well. Thus finding corresponding points is reduced to the solution of equation

$$\det[\mathbf{c}_{\mathbf{a}}(u), \dot{\mathbf{c}}_{\mathbf{a}}(u), \mathbf{a}, \mathbf{c}_{\mathbf{b}}(v)] = 0. \tag{7}$$

Rational parameterization $u = \phi(s)$, $v = \psi(s)$ of any component of (7) leads to a parameterization of ruled surface $\mathbf{c}_{\mathbf{a}}(\phi(s)) + t\mathbf{c}_{\mathbf{b}}(\psi(s))$, where $\mathcal{C}_{\mathbf{a}}$ is a contour w.r.t. $\mathbf{a}$. The second curve is a contour w.r.t. $\mathbf{b}$ iff $(\phi(s), \psi(s))$ is a parameterization of some component of curve

$$\det[\mathbf{c}_{\mathbf{b}}(u), \dot{\mathbf{c}}_{\mathbf{b}}(u), \mathbf{b}, \mathbf{c}_{\mathbf{a}}(v)] = 0, \tag{8}$$

by the same arguments. So let Δ denote the greatest common divisor of left hand sides of (7) and (8) and let $\Delta = \Delta_1 \cdots \Delta_k$ be a factorization to reducible components. Then any $\Delta_i(u, v)$ defining a rational curve in the space of parameters leads to desired ruled surfaces. With some additional work one can prove the theorem.

Theorem 3.2 *Two contour lines determine rational ruled surface uniquely.*

Remark 3.3 *Because of the form of reparameterizations $u = \phi(s)$ and $v = \psi(s)$, the unique rational component of Δ can be written as*

$$\gamma uv - \alpha u + \delta v - \beta = 0, \tag{9}$$

for some constants α, β, γ and δ.

Surprisingly, unlike the quadratic case, two silhouettes are enough to reconstruct any non-quadratic ruled surface.

Theorem 3.4 *Ruled surface of degree at least* 3 *is determined uniquely by two its generic silhouettes.*

3.2 Isophotes

Recall that we used a slightly modified definition of the isophote – in our case $\mathcal{I}_{\mathbf{a},\alpha}$ is a loci of points on the surface whose normal vector forms the angle $\pm\phi = \pm \arccos \alpha$ with direction $\mathbf{a}$. It turns out that in some cases $\mathcal{I}_{\mathbf{a},\alpha}$ consists of two components $\mathcal{I}_{\mathbf{a},\alpha} = \mathcal{I}^+_{\mathbf{a},\alpha} \cup \mathcal{I}^-_{\mathbf{a},\alpha}$ corresponding to the choice of the sign of the angle. There exists a close connection between existence of these decompositions and the offset of the surface. See [8, 2, 10] for definitions and algebraic/geometric properties of the offsets.

Lemma 3.5 *The isophotes on a ruled surface fall into components corresponding to angles $\pm\phi$ if and only if the surface has reducible offset.*

 ISBN: 978-84-608-6082-2

It can be seen easily that the both components $\mathcal{I}^{\pm}_{\mathbf{a},\alpha}$ are rational curves. If the isophote is irreducible then it doubly covers the projective line and the following theorem shows that even on regular quadrics the isophotes are not rational curves

Theorem 3.6 *The irreducible isophote is a hyperelliptic curve with genus less of equal to* $\deg \mathcal{R} - 1$. *Moreover if* $\mathcal{R}$ *does not contain singular ruling then the equality holds for almost all irreducible isophotes.*

Since a hyperelliptic curve of genus k posses at most $k+1$ real components (in a projective extension and after desingularization) we arrive at corollary.

Corollary 3.7 *The real part of irreducible* $\mathcal{I}_{\mathbf{a},\alpha}$ *consists of at most* $\deg \mathcal{R}$ *components.*

4 Conclusion

This paper was devoted to the study of contours and isophotes on ruled surfaces. We also presented the existence of the solution to the reconstruction problems. Although ruled surfaces are used a lot in applications, contours and isophotes on other commonly used surfaces are not rational. There might be two possible directions of further research. First one is the study of surfaces containing a lot of rational contour lines – for example on dual to Del Pezzo surface there exists at least two parametric family of contour curves. Secondly, we can leave the strong assumption on rationality and focus on surface with manageable contour curves – for example hyperelliptic curves posses relatively simple non-rational parameterizations, which allows to extend the class of studied surfaces e.g. by envelopes of quadratic cones and mentioned duals to Del Pezzo surfaces.

Acknowledgements

The author is supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

[1] M. Aigner, L. Gonzalez-Vega, B. Jüttler, and M. L. Sampoli, *Mathematics of Surfaces XIII: 13th IMA International Conference York, UK, September 7-9, 2009 Proceedings*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, ch. Computing Isophotes on Free-Form Surfaces Based on Support Function Approximation, pp. 1–18.

[2] E. Arrondo, J. Sendra, and J. R. Sendra, *Parametric generalized offsets to hypersurfaces*, Journal of Symbolic Computation, 23 (1997), pp. 267–285.

[3] M. Bizzarri and M. Lávička, *Parameterizing rational offset canal surfaces via rational contour curves*, Computer-Aided Design, 45 (2013), pp. 342–350.

[4] D. A. Cox, J. Little, and D. O'Shea, *Using algebraic geometry*, Springer-Verlag, 2 ed., 2005.

[5] Y. Y. Fatih Dogan, *On isophote curves and their characterizations*, Turkish Journal of Mathematics, 39 (2015), pp. 650–664.

[6] K.-J. Kim and I.-K. Lee, *Computing isophotos of surface of revolution and canal surface*, Computer Aided Design, 35 (2003), pp. 215–223.

[7] B. R. Martin Peternell, Helmut Pottmann, *On the computational geometry of ruled surfaces*, Computer-A, 31 (1999), pp. 17–32.

[8] H. Pottmann, *Rational curves and surfaces with rational offsets*, Computer Aided Geometric Design, 12 (1995), pp. 175–192.

[9] H. Theisel, *Are isophotes and reflection lines the same?*, Computer, 18 (2001), pp. 711–722.

[10] J. Vršek and M. Lávička, *Surfaces with pythagorean normals along rational curves*, Computer Aided Geometric Design, 31 (2014), pp. 451 – 463. Recent Trends in Theoretical and Applied Geometry.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Fast numerical valuation of European options under Merton's jump-diffusion model

Wansheng Wang[1] **and Yingzi Chen**[1]

[1] *School of Mathematics and Computational Science, Changsha University of Science & Technology, 410114, Hunan, China*

emails: w.s.wang@163.com, fenggegevip163@qq.com

Abstract

In this paper, we consider discontinuous Galerkin (DG) finite element together with finite difference (FD) scheme for solving Merton's jump-diffusion model, which is given by a partial integro-differential equations (PIDEs). Spatial differential operators are discretized using FD on a uniform grid, and time stepping is performed using the DG finite element method. The treatment of the integral term associated with jumps in models is more challenging. The discretization of this integral term will lead to full matrices for the non-locality of the integral operator. To fast solve this model, multigrid method is used for solving such linear algebraical system. Numerical results show that multigrid solver can speed up the existing DG-FD method. Some properties of this scheme are also analyzed.

Key words: Merton's jump-diffusion model, option pricing, PIDEs, discontinuous Galerkin finite element method, finite difference method, stability and convergence, positivity, multigrid,

1 Introduction

One of the modern financial theorys biggest successes in terms of both approach and applicability has been the BlackCScholes option pricing model developed by Fisher Black and Myron Scholes in 1973 [1] and previously by Robert Merton [2]. The celebrated Black-Scholes model is based on assumption that the price of the underlying asset behaves like a geometric Brownian motion with a drift and a constant volatility which cannot explain the market prices of options with various strike prices and maturities. To explain these behaviour, a number of alternative models have appeared in the financial literature. In

1976, Merton proposed to add to the behaviour of asset prices jumps which have normally distributed size. However, these model are more difficult to handle numerically in contrast to the celebrated Black-Scholes model. For example, for the Merton's jump-diffusive model, which is given by a partial integro-differential equations (PIDEs), the integral term in the PIDEs is on a semi-infinite interval. How are we going to approximate the integral term? Its discretization generally leads to a dense system. How to solve this dense system? This is our main purpose of this paper. To exploit the time analyticity of the slution for $t > 0$ and to compe with the loss of this analyticity at $t = 0$ [4], we consider discontinuous Galerkin (DG) time stepping scheme in combination with finite difference scheme for such Merton' model. Some properties of this scheme will be analyzed.

2 Merton's model and option pricing problems

Let $V(t, S)$ be the value of a European contract that depends on the time t and underlying asset price S, which is given by a process of the form

$$\frac{dS}{S} = \nu dt + \sigma dz + (\eta - 1)dq, \tag{1}$$

where ν is the drift rate, σ is the volatility of the Brownian part of the process, $\eta - 1$ is an impulse function giving a jump from S to $S\eta$, and dq is a Poisson process and assumed to be independent of the Wiener process dz. Here, $dq = 0$ with probability $1 - \lambda dt$, $dq = 1$ with probability λdt, where λ is the Poisson arrival intensity.

Under the above assumptions it is well known (Merton, 1976, [3]) that $V(t, S)$ satisfies a final value problem defined by the following PIDE:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - \lambda\kappa)S\frac{\partial V}{\partial S} - (r + \lambda)V + \lambda I(V(t, S)) = 0, \tag{2}$$

where r is the risk-free interest rate, κ denotes the average relative jump size, $\mathbb{E}(\eta - 1)$, and $I(V(t, S))$ denotes the integral $I(V(t, S)) = \int_0^\infty V(t, S\eta)g(\eta)d\eta$. Here $g(\eta)$ is the probability density function of the jump amplitude η; thus for all η, $g(\eta) \geq 0$, and $\int_0^\infty g(\eta)d\eta = 1$. In the model originally presented by Merton [3] the probality density function of the jump is

$$g(\eta) = \frac{1}{\sqrt{2\pi}\gamma\eta} e^{-[\ln\eta - \mu]^2/2\gamma^2} \tag{3}$$

The expected relative change in the stock price is $\kappa = \mathbb{E}(\eta - 1) = e^{\mu + \gamma^2/2} - 1$. The terminal condition is the following

$$V(T, S) = \begin{cases} (S - K)^+, & \text{in the case of a call option,} \\ (K - S)^+, & \text{in the case of a put option.} \end{cases} \tag{4}$$

 ISBN: 978-84-608-6082-2

The payoff function of (4) has a slop discontinuity at $S = K$. By making the changes of variables $x = \ln S$, $y = \ln \eta$ $(0 < \eta < \infty)$, $\tau = T - t$, $V(T - t, e^x) = u(\tau, x)$, evaluation of the option values requires solving the PIDE

$$\frac{\partial u}{\partial \tau} = \frac{1}{2}\sigma^2 \frac{\partial^2 u}{\partial x^2} + (r - \frac{1}{2}\sigma^2 - \lambda\kappa)\frac{\partial u}{\partial x} - (r + \lambda)u + \lambda \int_{-\infty}^{+\infty} u(\tau, x + y) f(y) dy = 0, \qquad (5)$$

with appropriate initial and boundary conditions, where $f(y) = \frac{1}{\sqrt{2\pi}\gamma} \exp\left[-\frac{(y-\mu)^2}{2\gamma^2}\right]$.

3 Finite difference spatial discretization

3.1 Finite difference discretization of spatial derivatives

Here we describe the discretization of the spatial derivative terms together with the zeroth-order term, that is, the operator

$$\mathcal{L}^{BS}[u] = \frac{1}{2}\sigma^2 \frac{\partial^2 u}{\partial x^2} + (r - \frac{1}{2}\sigma^2 - \lambda\kappa)\frac{\partial u}{\partial x} - (r + \lambda)u. \qquad (6)$$

A uniform mesh $-x^* = x_0 < \cdots < x_{i-1} < x_i < x_{i+1} < \cdots < x_M = x^*$ is used, where x^* will be appropriately chosen. The space derivatives of (6) are approximated with central-differences

$$\frac{\partial u}{\partial x}(\tau, x_i) \approx \frac{u_{i+1}(\tau) - u_{i-1}(\tau)}{2h}, \quad \frac{\partial^2 u}{\partial x^2}(\tau, x_i) \approx \frac{u_{i+1}(\tau) - 2u_i(\tau) + u_{i-1}(\tau)}{h^2}, \qquad (7)$$

where $h = x_{i+1} - x_i$, $u_i(\tau) = u(\tau, x_i)$. Then it follows that

$$\begin{aligned} \mathcal{L}_h^{BS}[u] \;=\; & \frac{1}{2}\sigma^2[\frac{u_{i+1}(\tau) - 2u_i(\tau) + u_{i-1}(\tau)}{h^2}] + (\nu - \frac{1}{2}\sigma^2 - \lambda\kappa)\frac{u_{i+1}(\tau) - u_{i-1}(\tau)}{2h} \\ & -(\nu + \lambda)u_i(\tau), \end{aligned} \qquad (8)$$

which will lead to a tridiagonal matrix which we denote by L_1.

3.2 Approximating integrals

The discretization of the integral operator

$$\mathcal{L}^{jump}[u] = -\lambda \int_{\mathbb{R}} u(\tau, x + y) f(y) dy. \qquad (9)$$

in (5) leads to a full matrix which we denote by L_2. By making the change of variable $y = z - x$, we obtain, $\int_{\mathbb{R}} u(\tau, x + y) f(y) dy = \int_{\mathbb{R}} u(\tau, z) f(z - x) dz$. Next, we split the

integral on the right-hand side as $\int_{\mathbb{R}} = \int_{\Omega_*} + \int_{\mathbb{R}\setminus\Omega_*}$, where $\Omega_* := (-x^*, x^*)$. On the interval $\Omega_* = (-x^*, x^*)$, we use the composite trapezoidal rule to approximate the integral and have

$$\begin{aligned}\mathcal{L}^{jump}[u] &= \Big(\int_{-\infty}^{-x^*} + \int_{x^*}^{+\infty} + \int_{-x^*}^{x^*}\Big)u(\tau, x_i + y)f(y)dy \\ &\approx e^{x_i+\frac{\gamma}{2}}\Phi\Big(\frac{x_i - x^* + \gamma^2}{\gamma}\Big) - Ke^{-r\tau}\Phi\Big(\frac{x_i - x^*}{\gamma}\Big) + \frac{h}{2}\Big[f_{i,0}u_0 + f_{i,M}u_M + 2\sum_{j=1}^{M-1} f_{i,j}u_j\Big], \quad (10)\end{aligned}$$

where $f_{ij} = f(x_j - x_i)$ and $\Phi(y) = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{y} e^{-\frac{t^2}{2}}dt$. Then we obtain a semi-discretize system

$$\frac{du}{d\tau} + Lu(\tau) = F(\tau), \quad \tau \in J = (0, T], \tag{11}$$

where $L = L_1 - h\lambda L_2$, L_1 is a tridiagonal matrix and L_2 is a full matrix.

4 Discontinuous Galerkin time discretization

We discretize (11) in time using a DG method following [4]. For $0 < T < \infty$ and $N \in \mathbb{N}$, let $N(\mathcal{N}) = \{I_k\}_{k=1}^N$ be a partition of $J = (0, T]$ into N subintervals $I_k = (\tau^{k-1}, \tau^k)$, $k = 1, 2, \cdots, N$, with $0 =: \tau^0 < \tau^1 < \cdots < \tau^{N-1} < \tau^N := T$. Moreover, denote by $\Delta\tau_k := \tau^k - \tau^{k-1}$ the length of I_k. Set $\Delta\tau = \max_{1\le k\le N}\Delta\tau^k$. Let $u \in L^2(I_k; X), u' \in L^2(I_k; X^*), w \in L^2(J, W), k = 1, 2, \ldots, N$, and define the one-sided limits

$$u_+^k = \lim_{s\to 0, s>0} u(\tau^k + s), \quad 0 \le k \le N-1; \quad u_-^k = \lim_{s\to 0, s>0} u(\tau^k - s), \quad 1 \le k \le N,$$

and the jumps

$$[u]_k := u_+^k - u_-^k, \qquad 1 \le k \le N-1.$$

To each time interval I_k, a polynomial degree (approximation order) $r_k \ge 0$ is associated. These numbers are stored in the degree vector $\underline{r} = \{r_k\}_{k=1}^N$. Then, the following space being used for the DG method is introduced:

$$\mathcal{S}^r(\mathcal{N}, W) := \{u \in L^2(J, W) : u|_{I_k} \in \mathcal{P}_{r_k}(I_k, W), k = 1, 2, \ldots, N\},$$

where $\mathcal{P}_{r_k}(I_k)$ denotes the space of polynomials of degree at most r_k on I_k. We define the bilinear form B_{dG} and the linear form F_{dG} by

$$B_{dG}(u, w) := \sum_{k=1}^{N}\int_{I_k}\{(u', w)_k + L(u, w)_k\}dt + \sum_{k=1}^{n-1}([u], w_+)_k + (u_+, w_+)_0, \tag{12}$$

$$F_{dG}(w) := \sum_{k=1}^{N}\int_{I_k}(F(\tau), w)dt + (u, w_+)_0. \tag{13}$$

 ISBN: 978-84-608-6082-2

Now the DG finite element method for (11) reads as follows: Find $U^{dG} \in \mathcal{S}^r_-(\mathcal{N}, W)$ such that

$$B_{dG}(U^{dG}, W) = F_{dG}(W), \quad \forall W \in \mathcal{S}^r_-(\mathcal{N}, W). \tag{14}$$

With the above discreteization, we obtain a linear system:

$$Q\upsilon^k = b^k, \quad k = 1, 2, \cdots, N, \tag{15}$$

where $Q = [\hat{A}]^k + \frac{\Delta\tau^k}{2}(I \otimes L)$ is a non-symmetric full matrix.

5 Properties of the numerical scheme

In this section, we investigate into some properties of the numerical scheme proposed here. We first consider the stability and positivity property. Since the problems model the hedge cost of replication of a contingent claim and the value of option is nonnegative, a nice property of the numerical scheme for the pricing equation is positivity-preserving. We also prove that this scheme is stable and convergent for Merton's model.

6 Multigrid method for algebraic system

How to solve the linear system with a dense matrix Q? In the literature, various methods including incomplete GMRES and iterative methods based on regular splittings are proposed. In this paper, we use geometric multigrid method to solve this algebraic system. Since the relaxation scheme used here is Gauss-Seidel iterative method, which generally requires that the matrix Q is positive definite, we should treat the algebraic system such that its coefficient matrix is symmetric positive definite. To do this, we solve the following equation with multigrid method

$$Q^T Q\upsilon^k = Q^T b^k, \quad k = 1, 2, \cdots, N.$$

7 Numerical examples

We present some numerical results for European call and put options under the Merton's jump diffusion model. Tables 1 and 2 below compare Merton's exact formula for European call and put with the prices generated by the DG-FD scheme together with multigrid solver.

European call option. Let the parameters in model be

$$\sigma = 0.15, \quad r = 0.05, \quad T = 0.25, \quad K = 100, \quad \lambda = 0.1, \quad \gamma = 0.5, \quad S_{\max} = 400.$$

The numerical results are presented in Table 1.

 ISBN: 978-84-608-6082-2

Table 1: Numerical results of DG-FD with multigrid solver for Merton's model

M	N	S	Computed values	Error	R	CPU time(s)
40	6	100	3.9069915924	0.0263548929		0.020752
80	10	100	3.8712451736	0.0093915259	2.83	0.075231
160	18	100	3.8831123893	0.0024756898	3.85	0.253685
320	34	100	3.8811509635	0.0005142641	4.87	1.475006
640	66	100	3.8807318342	0.0000951347	5.44	8.265404

European put option. Let the parameters in model be

$$\sigma = 0.2, \quad r = 0, \quad T = 0.2, \quad K = 1, \quad \lambda = 0.1, \quad \gamma = 0.5, \quad S_{\max} = 6.$$

The numerical results are presented in Table 2.

Table 2: Numerical results of DG-FD with multigrid solver for Merton's model

M	N	S	Computed values	Error	R	CPU time(s)
40	6	1	0.0317840160	0.0024715869		0.018392
80	10	1	0.0348437285	0.0005881255	4.20	0.069201
160	18	1	0.0343828579	0.0001272549	4.61	0.225363
320	34	1	0.0342296882	0.0000259148	4.89	1.342957
640	66	1	0.0342608248	0.0000052218	4.96	6.241253

Acknowledgements

This work was supported by the Natural Science Foundation of China (Grant No. 11371074) and the Natural Science Foundation for Distinguished Young scholars in Hunan Province, China (Grant No. 13JJ1020), and the Research Foundation of Education Bureau of Hunan Province, China (Grant No. 13A108).

References

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

[2] R. C. Merton, *Theory of rational option pricing*, Rand J. Econ., **4** (1973) 141–183.

[3] R. C. Merton, *Option pricing when underlying stock returns are discontinuous*, J. Financial Econ., **3** (1976) 125–144.

[4] A. M. Matache, C. Schwab and T. P. Wihler, *Fast numerial solution Of parabolic integro-differential equations with applications in finance*, IMA preprint series 1954, University of Minnesota, (2004).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Exact and discretized dissipativity of the nonliner functional-integro-differential equations

Liping Wen[1] **and Qing Liao**[1]

[1] *School of Mathematics and Computational Science, Xiangtan University, Xiangtan, Hunan, 411105, PR China*

emails: lpwen@xtu.edu.cn, qfashions@qq.com

Abstract

This paper is concerned with the dissipativity of a class of nonlinear functional-integro-differential equations (FIDEs). The dissipativity result of the theoretical solution for this class problem is presented. A type of extended one-leg methods are suggested for the FIDEs. It is shown under suitable condition that a $G(c, p, 0)$-algebraically stable one-leg methods is dissipative when they are used to the above problem. Numerical examples are given to illustrate the correctness of our theoretical results.

Key words: functional-integro-differential equations, one-leg method, dissipativity, algebraically stability, dynamical systems

1 introduction

Many interesting dynamical systems in physics and engineering are described by the property of a bounded absorbing set so that all trajectories enter in a finite time and thereafter remain inside (see [1]). Over the past few decades, the dissipativity results of the analytic solution and numerical methods for various type dynamical systems have been published. Among the authors of these papers we wish to mention just a few: R. Temam, A. R. Humphries, A. M. Stuart, A. T. Hill, C. M. Huang and S. Q. Gan. Early in 1994, Humphries and Stuart [2] studied the dissiativity of initial value problem and the Runge-Kutta methods for ordinary differential equations (ODEs) firstly. In 1997, Hill [3,4] obtained many results on the dissipativity of numerical methods for ODEs. In 2000, Huang [5] gave a sufficient condition for the dissipativity of theoretical solution of DDEs with constant delay, he also investigated the dissipativity of Runge-Kutta methods [5] and of one-leg

 ISBN: 978-84-608-6082-2

methos [6] for EEds. In 2006, Wen [7] have further discussed the dissipativity of the solution of general from Volterra functional differential equations (VFDEs) and the dissipativity results are given. The dissipativity of θ-methods and Multistep-Runge-Kutta methods for nonlinear Volterra delay-integro-differential equations (VDIDEs) are investigated by Gan [8] and Zhang [9]. Subsequently, Wen [10, 11] gave the dissipativity of the theoretical solutions and one-leg methods for neutral delay integro-differential equations. In this paper, we will study the dissipativity of a class of nonlinear functional-integro-differential equations (FIDEs) (see (1) in section 2). This paper is organized as follows. In section 2, the result on dissipativity of the theoretical solution are given. In section 3, the results of one-leg methods for FIDEs are obtained. In section 4, the numerical experiments are given to illustrate the theoretical results which we stated in previous sections.

2 The dissipativity of functional integro-differential equations

Let $\mathbb{C}^d$ be a d dimensional complex Euclidian space with the inner product $\langle\cdot,\cdot\rangle$ and the corresponding norm $\|\cdot\|$. For any given real symmetric positive definite $k\times k$ matrix $G=[g_{ij}]$, the norm $\|\cdot\|_G$ on $\mathbb{C}^{dk}:=(\mathbb{C}^d)^k$ is defined by

$$\|U\|_G=(\sum_{i,j=1}^{k} g_{ij}\langle u_i,u_j\rangle)^{\frac{1}{2}}, U=(u_1^T,u_2^T,\cdots,u_k^T)^T\in\mathbb{C}^{dk}.$$

Furthermore, for a symmetric matrix A, the symbols $\lambda_{\max}^A$ and $\lambda_{\min}^A$ denote the maximum and minimum eigenvalues of A respectively.

Consider the following nonlinear functional integro-differential equations(FIDEs) of the form (cf. [12, 13])

$$\begin{cases}\frac{d}{dt}\left[x(t)-\int_{t-\tau}^{t} g(t,\xi,x(\xi))d\xi\right]=f(t,x(t),x(t-\tau)), & t_0\le t<T\\ x(t)=\varphi(t), & t_0-\tau\le t\le t_0\end{cases}\tag{1}$$

where $T\le+\infty,\tau>0$ is a given constant delay, the functions $f:[t_0,T)\times\mathbb{C}^d\times\mathbb{C}^d\to\mathbb{C}^d$, $g:\mathbb{D}\times\mathbb{C}^d\to\mathbb{C}^d$, and $\varphi:[t_0-\tau,t_0]\to\mathbb{C}^d$ are assumed to be continuous so that system (1) has a unique solution $x(t)$ and satisfies the conditions

$$Re\langle f(t,u,v),u-w\rangle\le\gamma+\alpha\parallel u\parallel^2+\beta\parallel v\parallel^2+\eta\parallel w\parallel^2,\ t\ge t_0,\ u,v,w\in\mathbb{C}^d\tag{2}$$

$$\parallel g(t,\xi,u)\parallel\le\lambda\parallel u\parallel,\ (t,\xi)\in\mathbb{D},\ u\in\mathbb{C}^d\tag{3}$$

where $\lambda>0$, with $2\lambda^2\tau^2<1$; $\mathbb{D}:\{(t,\xi):t\in[t_0,T),\xi\in[t-\tau,t]\}$ and $\gamma,-\alpha,\beta,\eta$ are nonnegative constant. In order to investigate the numerical dissipativity of (1), we assume further that f holds the condition: for any constant $M>0$, there exists $L>0$

which is only dependent on M such that $\| f(t,u,v) \| \leq L$ holds for any $t \geq 0$ and $\| u \| \leq M, \| v \| \leq M, \| w \| \leq M$.

Definition 2.1.(cf. [11]) The problem (1) in FIDEs is said to be dissipative in $\mathbb{C}^d$ if there exists a bounded set $B \subset \mathbb{C}^d$, such that for any given bounded set $\Phi \subset \mathbb{C}^d$, there is a time $t^* = t^*(\Phi)$ such that for any given continuous initial function $\varphi : [t_0 - \tau, t_0] \to \mathbb{C}^d$ with $\varphi(t)$ contained in Φ for all $t \in [t_0 - \tau, t_0]$, the values of the corresponding solution $x(t)$ of the problem are contained in B for all $t \geq t^*$. Here B is called an absorbing set of the problem.

In order to study the dissipativity of (1), the following lemma play a key role:

Lemma 2.2 (see [11]) If $u(t)$, $w(t) \geq 0$, $t \in [t_0 - \tau, +\infty]$,

$$\begin{cases} u^{'}(t) \leq R(t) + A(t)u(t) + B(t) \sup\limits_{t-\tau \leq \xi \leq t} w(\xi), \quad t \geq t_0, \\ w(t) \leq G(t)u(t) + H(t) \sup\limits_{t-\tau \leq \xi \leq t} w(\xi) \end{cases} \tag{4}$$

and

$$\sup_{t_0-\tau \leq \xi \leq t_0} w(\xi) \leq \frac{G_0}{1 - H_0} \sup_{t_0-\tau \leq \xi \leq t_0} u(\xi), \tag{5}$$

where $A(t)$ is a continuous function satisfying $A(t) \leq A_0$ with constant $A_0 < 0$; $R(t), B(t), G(t)$ and $H(t)$ are nonnegative continuous functions satisfying $G(t) \leq G_0$, $H(t) \leq H_0$ with constants $G_0 \geq 0$, $0 \leq H_0 < 1$, for $t \in [t_0, +\infty]$; $\tau \geq 0$ is a constant, and if there exists $0 \leq p < 1$ such that

$$pA(t) + \frac{G_0}{1 - H_0} B(t) \leq 0, for\ t \geq t_0, \tag{6}$$

holds, then for $t \geq t_0$ we have

$$\begin{cases} u(t) \leq \frac{-\gamma^*}{(1-p)A_0} + \phi e^{-\mu^*(t-t_0)}, \\ w(t) \leq \frac{G_0}{1-H_0} \frac{-\gamma^*}{(1-p)A_0} + \frac{G_0}{1-H_0 e^{\mu^* \tau}} \phi e^{-\mu^*(t-t_0)} \end{cases} \tag{7}$$

where

$$\phi = \sup_{t_0-\tau \leq \xi \leq t_0} u(\xi), \quad \gamma^* = \sup_{t_0 \leq t < +\infty} R(t) \tag{8}$$

and $\mu^* > 0$ is defined as

$$\mu^* = \inf_{t \geq t_0} \left\{ \mu(t) : \mu(t) + A(t) + B(t) \frac{G_0 e^{\mu(t)\tau}}{1 - H_0 e^{\mu(t)\tau}} = 0 \right\}. \tag{9}$$

 ISBN: 978-84-608-6082-2

By applying **Lemma 2.2** we can prove the following theorem.

Theorem 2.3 Suppose that $x(t)$ is a solution of the problem (1) where f and g satisfy (2) with $\alpha < 0$ and (3), respectively, and there exists constant $0 < p < 1$ such that

$$p\alpha + \frac{4}{1-2\lambda^2\tau^2}(\beta^+ + (\eta-\alpha)\lambda^2\tau^2) \leq 0,\ t \geq t_0 \tag{10}$$

then,
(i) for any $t \geq 0$, we have

$$\| x(t) \|^2 \leq \frac{4}{1-2\lambda^2\tau^2}\frac{-\gamma^*}{(1-p)\alpha} + \frac{1-2\lambda^2\tau^2}{1-2\lambda^2\tau^2 e^{\mu^*\tau}}\phi e^{-\mu^*(t-t_0)},$$

where $\phi = \underset{t_0-\tau\leq\xi\leq t_0}{sup} \| \varphi(\xi) \|^2$, and $\mu^* > 0$ is defined as

$$\mu^* = \inf_{t\geq t_0}\left\{\mu(t) : \mu(t) + \alpha + (\beta^+ + (\eta-\alpha)\lambda^2\tau^2)\frac{4e^{\mu(t)\tau}}{1-2\lambda^2\tau^2 e^{\mu(t)\tau}} = 0\right\};$$

(ii) for any given $\varepsilon > 0$, the problem (1) is dissipativity with an absorbing set

$$B = B\left(0, \sqrt{\frac{4}{1-2\lambda^2\tau^2}\frac{-\gamma^*}{(1-p)\alpha} + \varepsilon}\right),$$

here and later, we denote that $\beta^+ = \max\{\beta, 0\}$.

3 Dissipativity of one-leg methods

As we all know that one-leg methods are a class of effective methods for OEDs, which can be expressed as

$$\rho(E)x_n = hf(\sigma(E)t_n, \sigma(E)x_n) \tag{11}$$

in which $h > 0$ is the computational step size, x_n is an approximation to $x(t_n)$, $t_n = t_0 + nh$, E denotes the shift operator:$Ex_n = x_{n+1}$, and the polynomials

$$\rho(\zeta) = \sum_{j=0}^{k}\alpha_j\zeta^j,\ \sigma(\zeta) = \sum_{j=0}^{k}\beta_j\zeta^j,$$

are assumed to have real coefficients with $\alpha_k \neq 0$, $\alpha_0 + \beta_0 \neq 0$, no common divisor and satisfy the consistent conditions:

$$\rho(1) = 0,\ \rho'(1) = \sigma(1) = 1.$$

Adapting methods (11) to systems (1) leads to

$$\begin{cases} \rho(E)(x_n - z_n) = hf(\sigma(E)t_n, \sigma(E)x_n, \sigma(E)x_{n-m}), \quad n \geq 0 \\ z_n = h\sum\limits_{i=0}^{m} v_i g(t_n, t_{n-i}, x_{n-i}) \end{cases} \tag{12}$$

where $t_n = nh$, step size $h = \frac{\tau}{m}$, m is a given positive integer, x_n is approximations to $x(t_n)$ and z_n is an approximations to $z(t_n) := \int_{t_{n-\tau}}^{t_n} g(t_n, \xi, x(\xi))d\xi$ respectively. In methods (12), we usually adopt the compound quadrature rules to discretize the integral items. In the following, we always assume that the compound quadrature rules which we used satisfy (cf. [12, 12])

$$h\sqrt{(m+1)\sum_{i=0}^{m}|v_i|^2} < v \tag{13}$$

where $v > 0$ is a constant with $4v^2\lambda^2 < 1$. Here λ is given in (3).

In addition, for any given sufficiently smooth function $y(t)$, we have

$$\sigma(E)y(t_n) = y(\sigma(E)t_n) + O(h^2).$$

Therefore, we will further assume that the quadrature rules satisfies

$$\|\sigma(E)z_n\| \leq h\|\sum_{i=0}^{m} v_i g(\sigma(E)t_n, \sigma(E)t_{n-i}, \sigma(E)x_{n-i})\|. \tag{14}$$

This section will focus on the dissipativity analysis of $G(c, p, 0)$-algebraically stable one-leg methods with respect to nonlinear FIDEs (1). Now we introduce some useful definitions and lemma as follows firstly.

Definition 3.1. (See [14]). Let G be a real $k \times k$ symmetric positive definite matrix. An one-leg method is said to be $G(c, p, q)$-algebraically stable, if for all real $a_0, a_1, \ldots, a_k$,

$$A_1^T G A_1 - cA_0^T G A_0 \leq 2\sigma(E)a_0\rho(E)a_0 - p(\sigma(E)a_0)^2 - q(\rho(E)a_0)^2,$$

where $A_i = (a_i, a_{i+1}, \ldots, a_{i+k-1})^T, i = 0, 1$. As an important special case, a $G(1, 0, 0)$-algebraically stable method is called G-stable for short.

Lemma 3.2. (See [6]). Suppose $\{\xi_i(x)\}_{i=1}^{q}$ are a basis of polynomials for P^{q-1}, the space of polynomials of degree strictly less than q. Then, there is always a unique solution $y_n, \ldots, y_{n+q-1}$ to the system of equations

$$\xi_i(E)y_n = \Delta_i, \ \Delta_i \in \mathbb{C}^d, \ i = 1, \ldots, q,$$

 ISBN: 978-84-608-6082-2

and there exist a constant ς, independent of the Δ_i, such that

$$\max_{0\le i\le q-1}\|y_{n+i}\| \le \varsigma \max_{1\le i\le q}\|\Delta_i\|.$$

Definition 3.3. A method (12) is said to be dissipative if, whenever the method is applied with a step size h to a dynamical system of the form (1) subject to (2) and (3), there exists a constant r such that, for any function $\varphi(t)$, there exists an n_0 dependent only on $\varphi(t)$ and initial values $y_0, y_1, \ldots, y_{k-1}$, such that

$$\|y_n\| \le r,\ n \ge n_0$$

holds.

Now the dissipativity of one-leg methods (12) is state as follows.

Theorem 3.4. Assume that the one-leg method (11) is $G(c,p,0)$-algebraically stable with $c \le 1$ and that the problem (1) satisfies (2), (3) and (10). Then when $h(\alpha+\beta^+ + \eta v^2\lambda^2) < p^-(1+v^2\lambda^2)$, the method (12) for FIDEs (1) is dissipative, where $p^- = \min\{p,0\}$.

Corollary 3.5 Assume that the one-leg method (11) is A-stable and that the problem (1) satisfies (2),(3) and (10). Then when $\alpha+\beta^+ + \eta v^2\lambda^2 < 0$, the method (12) for FIDEs (1) is dissipative.

4 Numerical experiment

As an example, we consider the following two-dimensional system

$$\begin{cases} \frac{d}{dt}\big(x_1(t) - \frac{1}{4\pi}\int_{t-\frac{\pi}{12}}^{t} e^{\xi-t}\big(7x_1(\xi)+3x_2(\xi)\big)d\xi\big) = -x_1(t) \\ \qquad\qquad + \frac{1}{96}\big(\overline{x_1}(t-\frac{\pi}{12}) + \sqrt{5}\overline{x_2}(t-\frac{\pi}{12})\big) + f_1(t), \\ \frac{d}{dt}\big(x_2(t) - \frac{1}{4\pi}\int_{t-\frac{\pi}{12}}^{t} e^{\xi-t}\big(3x_1(\xi)-x_2(\xi)\big)d\xi\big) = -x_2(t) \\ \qquad\qquad + \frac{1}{96}\big(\sqrt{5}\overline{x_1}(t-\frac{\pi}{12}) - 3\overline{x_2}(t-\frac{\pi}{12})\big) + f_2(t), \end{cases} \quad t \ge 0. \tag{15}$$

where a, b are any given constants, and

$$f_1(t) = cos(at) - a\ sin(at),\ \ f_2(t) = sin(bt) + b\ cos(bt),$$

$$\overline{x_1}(t-\frac{\pi}{12}) = \frac{x_1(t-\frac{\pi}{12})}{1+x_1^2(t-\frac{\pi}{12})},\ \overline{x_2}(t-\frac{\pi}{12}) = \frac{x_2(t-\frac{\pi}{12})}{1+x_2^2(t-\frac{\pi}{12})}.$$

For this system we have $2\tau^2\lambda^2 = \frac{1}{18} < 1$. We choose

$$\alpha = -\frac{23}{48}, \beta = \frac{1}{24}, \eta = \frac{25}{48}, \lambda = \frac{2}{\pi}, p = \frac{16}{25}, \gamma^* = 2\sqrt{(1-a)^2+(1+b)^2}, \tau = \frac{\pi}{12}.$$

Then all conditions of Theorem 2.4 are satisfied. Therefore, the system (15) is dissipative and $B = B(0, 7\sqrt{(1-a)^2 + (1+b)^2})$ is an absorbing set.

We use the second order BDF method

$$x_{n+2} - \frac{4}{3}x_{n+1} + \frac{1}{3}x_n = \frac{2}{3}hf(t_{n+2}, x_{n+2}) \tag{16}$$

with the composite trapezoidal rule

$$\int_{t_1}^{t_2} \phi(x)dx \cong h\Big[\frac{1}{2}\phi(t_1) + \sum_{j=1}^{m-1} \phi(t_1 + jh) + \frac{1}{2}\phi(t_2)\Big]$$

to solve the problem (15) and we select $v = \tau = \frac{\pi}{12}$ which satisfies the condition (13). The composite trapezoidal rule and method (16) have the same order(second order) so there is no order reduction. The method (16) is A-stable, so it is equivalent to G-stable, according to Corollary 3.5 the numerical solution is dissipative.

Let step size $h = 0.004\pi/12$, we consider different initial functions for $t \in [\frac{\pi}{12}, 0]$ as follows

(I) $y_1(t) = 2\ sin(t)e^t,\ y_2(t) = (t+1)^2 - 1$;
(II) $y_1(t) = sin(2t),\ y_2(t) = 2\ cos(3t)$;
(III) $y_1(t) = 1.5\ cos(5t),\ y_2(t) = sin(4t)$.

respectively, the numerical results are shown in Fig.1-3. These numerical examples prove that the problem (15) is dissipative. Therefore, this numerical example illustrate the correctness of our theoretical results.

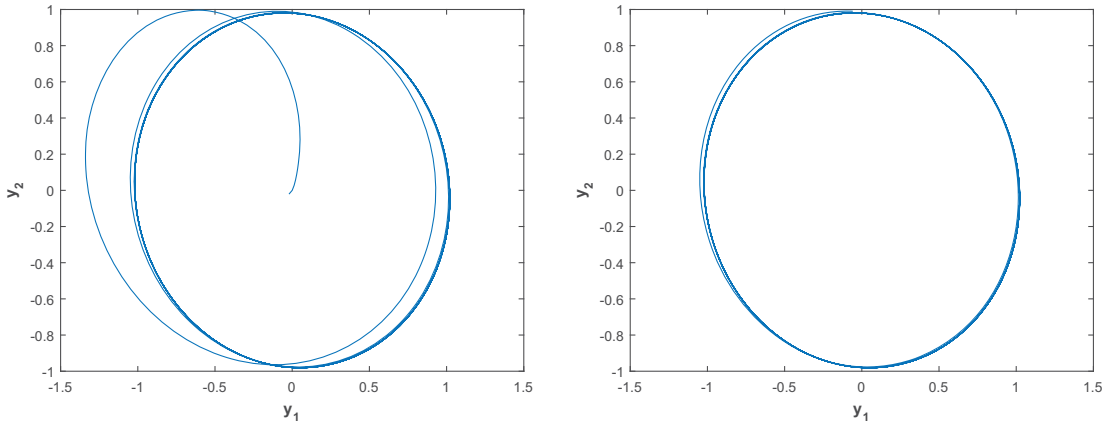


Figure 1: The numerical solution of (15) with initial function (I) and $a = 3,\ b = 3$ for $t \in [0, 10\pi]$ (left) and $t \in [\frac{5\pi}{6}, 10\pi]$ (right) respectively.

 ISBN: 978-84-608-6082-2

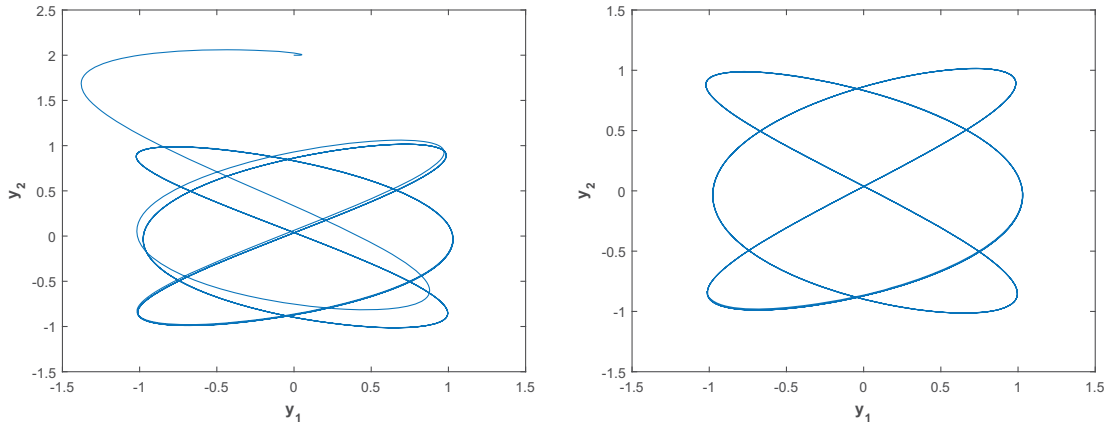


Figure 2: The numerical solution of (15) with initial function (II) and $a = 3,\ b = 2$ for $t \in [0, 10\pi]$ (left) and $t \in [\frac{5\pi}{3}, 10\pi]$ (right) respectively.

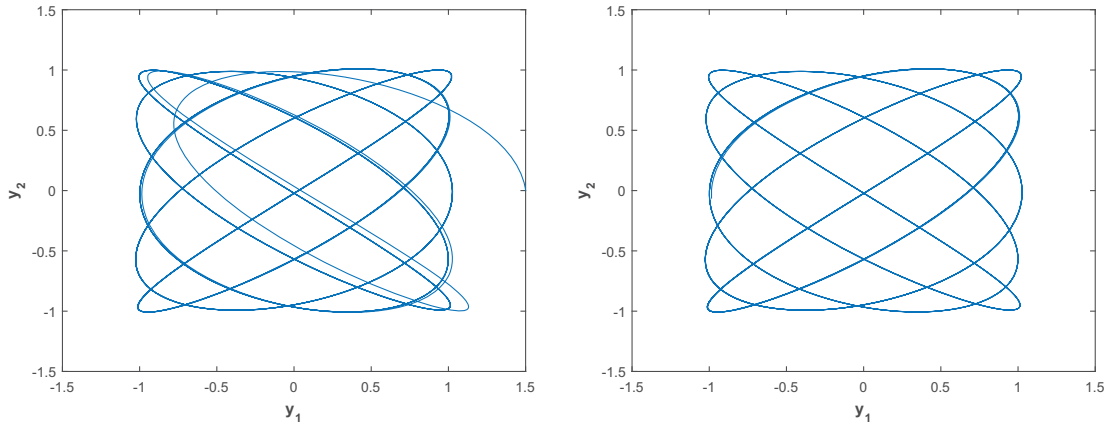


Figure 3: The numerical solution of (15) with initial function (III) and $a = 5,\ b = 4$ for $t \in [0, 10\pi]$ (left) and $t \in [\pi, 10\pi]$ (right) respectively.

Acknowledgements

This work was supported by the Natural Science Foundation of China (Grant No. 11371302).

References

[1] R. Temam, *Infinite-dimensional dynamical systems in mechanics and physics*, Springer Applied Mathematical Sciences Series , **68** (1988) Springer, Berlin.

[2] A.R. Humphries, A.M. Stuart, *Runge-Kutta methods for dissipative and gradient dynamical systems*, SIAM J. Numer. Anal. , **31** (1994) 1452–1485.

[3] A.T. Hill, *Global dissipativity for a-stable methods*, SIAM J. Numer. Anal. , **34** (1997) 119–142.

[4] A.T. Hill, *Dissipativity of Runge-Kutta methods in Hilbert spaces*, BIT , **37** (1997) 37–42.

[5] C.M. Huang, *Dissipativity of Runge-Kutta methods for dynamical systems with delays*, IMA J. Numer. Anal. , **20** (2000) 153–166.

[6] C.M. Huang, *Dissipativity of one-lag methods for dynamical systems with delays*, Appl. Numer. Math. , **35** (2000) 11–22.

[7] L.P. Wen, S.F. Li, *Dissipativity of Volterra functional differential equations*, J. Math. Anal. Appl. , **324** (2006) 696–706.

[8] S.Q. Gan, *Dissipativity of θ-methods for nonlinear Volterra delay-integro-differential equations*, J. Comput. Appl. Math. , **206** (2007) 898–907.

[9] R. Qi, C. Zhang, Y. Zhang, *Dissipativity of Multistep Runge-Kutta Methods for Nonlinear Volterra Delay-integro-differential Equations*, Acta Math. Appl. Sinica(English Series) , **28** (2012) 225–236.

[10] X.Y. Liu, L.P. Wen, *Dissipativity of one-leg methods for neutral delay integro-differential equations*, J. Comput. Appl. Math. , **235** (2010) 165–173.

[11] L.P. Wen, W.S. Wang, Y.X. Yu, *Dissipativity and asymptotic stability of nonlinear neutral delay integro-differential equations*, Nonlinear Anal.: TMA , **72** (2010) 1746–1754.

[12] C. Zhang, T. Qin, *The mixed Runge-Kutta methods for a class of nonlinear functional-integro-differential equations*, Appl. Math. Comput. , **237** (2014) 396–404.

[13] T. Qin, C. Zhang, *Stable solutions of one-leg methods for a class of nonlinear functional-integro-differential equations*, Appl. Math. Comput. , **250** (2015) 47–57.

[14] S. Li, *Theory of Computational Methods for Stiff Differential Equation*, Hunan Press of Science and Technology, Changsha, 1997.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

Quantum Wigner molecules in semiconductor quantum dots and cold-atom optical traps and their mathematical symmetries

Constantine Yannouleas[1] **and Uzi Landman**[1]

[1] *School of Physics, Georgia Institute of Technology, Atlanta, GA 30332-0430, USA*

emails: Constantine.Yannouleas@physics.gatech.edu,
Uzi.Landman@physics.gatech.edu

Abstract

Strong repelling interactions between a few fermions or bosons confined in two-dimensional circular traps lead to particle localization and formation of quantum Wigner molecules (QWMs) possessing definite point-group space symmetries. These point-group symmetries are "hidden" (or emergent), namely they cannot be traced in the circular single-particle densities (SPDs) associated with the exact many-body wave functions, but they are manifested as characteristic signatures in the ro-vibrational spectra. An example, among many, are the few-body QWM states under a high magnetic field or at fast rotation, which are precursor states for the fractional quantum Hall effect. The hidden geometric symmetries can be directly revealed by using spin-resolved conditional probability distributions, which are extracted from configuration-interaction (CI), exact-diagonalization wave functions. The hidden symmetries can also be revealed in the CI SPDs by reducing the symmetry of the trap (from circular to elliptic to quasi-linear). In addition the hidden symmetries are directly connected to the explicitly broken-symmetry (BS) solutions of mean-field approaches, such as unrestricted Hartree-Fock (UHF). A companion step of restoration of the broken symmetries via projection operators applied on the BS-UHF solutions produces wave functions directly comparable to the CI ones, and sheds further light into the role played by the emergence of hidden symmetries in the exact many-body wave functions. Illustrative examples of the importance of hidden symmetries in the many-body problem of few electrons in semiconductor quantum dots and of few ultracold atoms in optical traps (where unprecedented control of the interparticle interaction has been experimentally achieved recently) will be presented.

Key words: Wigner molecule, emergent point-group symmetries, broken symmetries, symmetry restoration, projection operator, unrestricted Hartree Fock, configuration interaction, 2D semicoductor quantum dots, trapped ultracold atoms, fermions, bosons

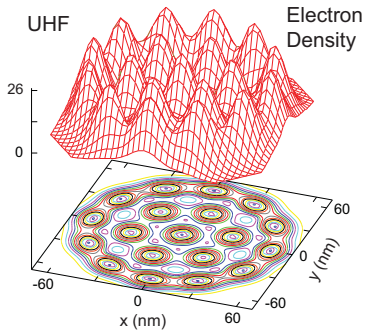


Figure 1: Unrestricted Hartree-Fock electron density in a 2D parabolic QD for $N = 19$ electrons and total-spin projection $S_z = 19/2$, exhibiting breaking of the circular symmetry at $R_W = 5$ and zero magnetic field. The electrons are (partially) localized in a (1,6,12) multi-ring structure. which exhibits point-group symmetries. Remaining parameters are: parabolic confinement, $\hbar\omega_0 = 5$ meV; effective mass $m^* = 0.067m_e$. Distances are in nanometers and the electron density in 10^{-4} nm^{-2}.

1 Introduction

This talk focuses on novel, somewhat exotic, types of clusters of few fermions or bosons. In particular, we discuss clusters of electrons in manmade (artificial) quantum dots (QDs) created through lithographic and gate-voltage techniques at semiconductor interfaces, and clusters of neutral ultracold atoms (either bosonic or fermionic) in harmonic optical traps. These cluster systems exhibit interesting emergent physical behavior arising from spontaneous breaking of spatial and/or spin symmetries at the *mean-field* level of theoretical treatment [1, 2]; symmetry breaking (SB) is defined as a circumstance where a lower energy solution of the Schrödinger equation is found that is characterized by a lower symmetry than that of the full many-body Hamiltonian of the few-body system. Such SB in circular traps directly reflects the localization of of particles in cluster arrangements exhibiting point-group symmetries instead of the continuous rotational symmetry expected from the many-body Hamiltonian [2]. A prominent example is the formation of finite electron crystallites (referred to as *semi-classical* Wigner molecules, SCWMs) in two-dimensional (2D) QDs (see Fig. 1). Symmetry breaking at the mean-field level is also manifested in the transition [3, 4], induced by increasing the interatomic repulsive contact-interaction strength, of the ground state of neutral atoms in a parabolic or toroidal 2D trap to a rotating bosonic quantum Wigner molecule (QWM). An example is presented in Fig. 2, where the hierarchy of the successive approximations (broken symmetry UHF $\rightarrow$ symmetry restoration) is illustrated, leading to the symmetry-restored, fully-quantal wave function (QWM) in Fig. 2 (c,d); the single-particle density (SPD) for the intermediate BS-UHF (SCWM) wave func-

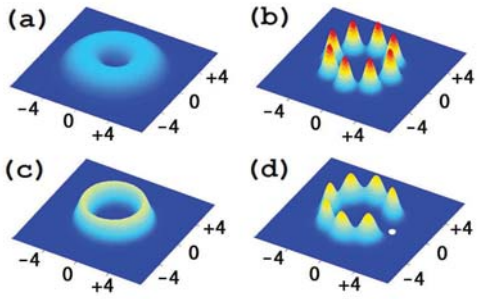


Figure 2: Single-particle densities and CPDs for $N = 8$ neutral repelling bosons in a rotating 2D toroidal trap with reduced rotational frequency $\Omega/\omega_0 = 0.2$ and $R_\delta = 50$. The confining potential, $m\omega_0^2(r - r_0)^2/2$, is centered at a radius $r_0 = 3l_0$. (a) Gross-Pitaevski SPD. (b) UHF SPD exhibiting breaking of the circular symmetry. (c) QWM SPD exhibiting circular symmetry. (d) CPD for the QWM wave function (resulting from the method of symmetry restoration), revealing the hidden point-group symmetry in the intrinsic frame of reference. The fixed observation point is denoted by a white dot. The QWM ground-state angular momentum is $L_z = 16$. Lengths in units of the oscillator length l_0. The vertical scale is the same for (b), (c), and (d), but different for (a).

tion is plotted in Fig. 2(b). Note that the point-group symmetry is not visible in the SPD after the step of symmetry restoration is executed, i.e., it becomes hidden [see Fig. 2(c)], but it is revealed via a conditional probability distribution (CPD); see Fig. 2(d).

The CPD gives the probability of finding a particle with spin σ at position $\mathbf{r}$ given that another one (referred to as the fixed particle) with spin σ_0 is located at $\mathbf{r}_0$. The degree of particle localization is controlled by the Wigner parameter that specifies the strength of the interparticle repulsion relative to the zero-point kinetic energy, i.e., $R_W = Z^2e^2/(\kappa l_0 \hbar\omega_0)$ [2, 3] for a Coulomb repulsion and $R_\delta = gm/(2\pi\hbar^2)$ [2, 3, 4] for a contact-potential (Dirac-delta) repulsion; Z is the charge of the particle, κ is the dielectric constant, ω_0 is the frequency of the harmonic trap, $l_0 = \sqrt{\hbar/(m\omega_0)}$ is the oscillator length, m is the particle mass, and g is the strength of the contact interaction.

Of great value in analyzing the physics associated with the hidden symmetries is the evolution of the lowest-energy band in the energy spectra (yrast band) as a function of the two successive approximations (broken symmetry UHF $\rightarrow$ symmetry restoration). Figs. 3(a,b) present the evolution of yrast spectra (as a function of the rotational frequency Ω of the trap) that are associated with the class of wave functions portrayed in Fig. 2. The most prominent trend is that the ground-state angular momenta of the symmetry-restored wave functions do not assume all the possible 2D values, but are restricted to stepwise values $L_z = Nk$, $k = 0, 1, 2, \ldots$, with $N = 8$, i.e., they change in steps of $N = 8$, where N is

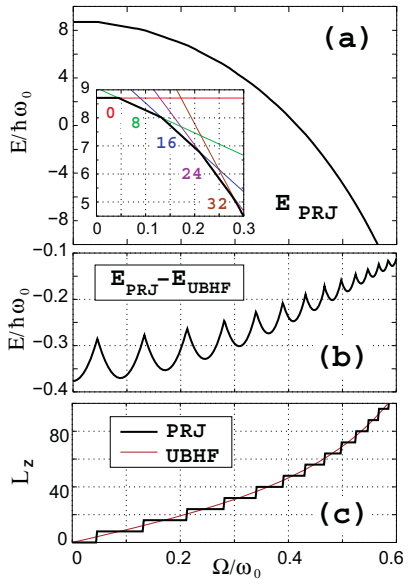


Figure 3: Properties of $N = 8$ neutral repelling bosons in a rotating 2D toroidal trap as a function of the reduced rotational frequency Ω/ω_0. The confining toroidal potential is centered at a radius $r_0 = 3l_0$, and the interaction-strength parameter was chosen as $R_\delta = 50$. (a) QWM ground-state energies, E^{PRJ}. The term QWM corresponds to projected (PRJ) wave functions that preserve the total angular momentum (symmetry restoration). The inset shows the range $0 \leq \Omega/\omega_0 \leq 0.3$. The numbers denote ground-state magic angular momenta. (b) Energy difference $E^{PRJ} - E^{UBHF}$, where the subscript "UBHF" stands for unrestricted bosonic Hartree-Fock. (c) Total angular momenta associated with (i) the QWM ground states [thick solid line (showing steps and marked as PRJ); online black] and (ii) the broken-symmetry UBHF solutions (smooth thin solid line; online red).

the number of particles [see Fig. 3(c) and the inset in Fig. 3(a)]. Such stepwise angular momenta are usually referred to as "magic" and the associated ground states of enhanced stability [see Fig. 3(b)] are finite-size precursors of the bulk fractional quantum Hall states; see also Section 3 below.

2 Group theoretical analysis of symmetry breaking in unrestricted Hartree Fock

We mention here the case of $N = 3$ fully spin polarized ($S_z = 3/2$) electrons in the absence of a magnetic field (B) and for $R_W = 10$ ($\kappa = 1.9095$). Fully spin polarized UHF determinants

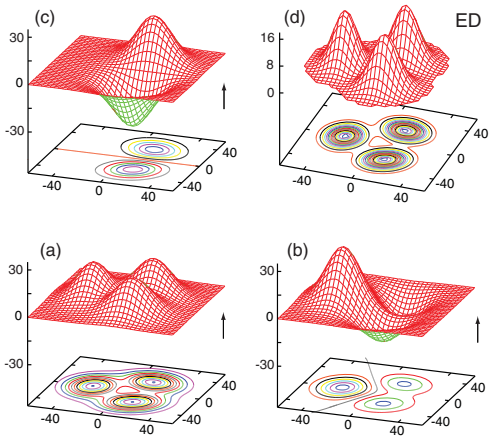


Figure 4: The UHF case exhibiting breaking of the circular symmetry for $N = 3$ electrons and total spin projection $S_z = 3/2$ at $R_W = 10$ and at zero magnetic field. (a-c): real orbitals (modulus square). (d): the corresponding electron density (ED). The choice of the remaining parameters is: $\hbar\omega_0 = 5$ meV and effective mass $m^* = 0.067m_e$. Distances are in nanometers. The real orbitals are in 10^{-3} nm^{-1} and the total ED in 10^{-4} nm^{-2}. The arrows indicate the spin projection ($S_z = 1/2$) for each orbital.

preserve the total spin, but for this value of R_W the lowest in energy UHF solution is one with broken circular symmetry. As it has been mentioned earlier, broken rotational symmetry does not imply no space symmetry, but a lower point-group symmetry [5].

In Fig. 4 we display the UHF symmetry-violating orbitals (a−c) whose energies are (a) 44.801 meV, (b) and (c) 46.546 meV, namely the two orbitals (b) and (c) with the higher energies are degenerate in energy. Overall the BS orbitals (a−c) drastically differ from the orbitals of the independent particle model. In particular, they are associated with specific sites (within the QD) forming an equilateral triangle, and thus they can be described as having the structure of a linear combination of "atomic" (site) orbitals (LCAOs). Such LCAO molecular orbitals (MOs) are familiar in natural molecules, and this analogy supports the term "semi-classical electron (or Wigner) molecules" for characterizing the BS-UHF solutions. We notice here that the LCAO orbitals in Fig. 4 are familiar in Organic Chemistry and are associated with the theoretical description of Carbocyclic Systems, and in particular the molecule C_3H_3 (cyclopropenyl, see, e.g., Ref. [6]). The important point of course is not the uniqueness or not of the 2D UHF orbitals, but the fact that they transform according to the irreducible representations of specific point groups, leaving both the UHF determinant and the associated electron densities invariant.

The electron density (ED) portrayed in Fig. 4(d) remains invariant under certain geo-

 ISBN: 978-84-608-6082-2

Table 1: Character table for the cyclic group C_3 [$\varepsilon = \exp(2\pi i/3)$]

C_3	E	C_3	C_3^2
A	1	1	1
E'	1	ε	ε^*
E''	1	ε^*	ε

metrical symmetry operations, namely those of an unmarked, plane and equilateral triangle. They are: (1) The identity E; (2) The two rotations C_3 (rotation by $2\pi/3$) and C_3^2 (rotation by $4\pi/3$); and (III) The three reflections σ_v^I, σ_v^{II}, and σ_v^{III} through the three vertical planes, one passing through each vertex of the triangle. These symmetry operations for the unmarked equilateral triangle constitute the elements of the group C_{3v} [6, 7].

One of the main applications of group theory in Chemistry is the determination of the eigenfunctions of the Schrödinger equation by simply using symmetry arguments alone. This is achieved by constructing the so-called *symmetry-adapted linear combinations* (SALCs) of AOs. A widely used tool for constructing SALCs is the projection operator

$$\hat{\mathcal{P}}^{\mu} = \frac{n_{\mu}}{|\mathcal{G}|} \sum_{R} \chi^{\mu}(R) \hat{R}, \tag{1}$$

where $\hat{R}$ stands for any one of the symmetry operations of the molecule, and $\chi^{\mu}(R)$ are the characters of the μth irreducible representation of the set of $\hat{R}$'s. (The χ^{μ}'s are tabulated in the socalled character tables [6, 7].) $|\mathcal{G}|$ denotes the order of the group and n_{μ} the dimension of the representation.

The task of finding the SALCs for a set of three $1s$-type AOs exhibiting the C_{3v} symmetry of an equilateral triangle can be simplified, since the pure rotational symmetry by itself (the rotations C_3 and C_3^2, and not the reflections σ_v's through the vertical planes) is sufficient for their determination. Thus one needs to consider the simpler character table of the cyclic group C_3 (see Table 1).

From Table 1, one sees that the set of the three $1s$ AOs situated at the vertices of an equilateral triangle spans the two irreducible representations A and E, the latter one consisting of two associted one-dimensional representations. To construct the SALCs, one simply applies the three projection operators $\hat{\mathcal{P}}^{A}$, $\hat{\mathcal{P}}^{E'}$, and $\hat{\mathcal{P}}^{E''}$ to one of the original AOs, let's say the ϕ_1,

$$\begin{aligned} \hat{\mathcal{P}}^{A}\phi_1 &\approx (1)\hat{E}\phi_1 + (1)\hat{C}_3\phi_1 + (1)\hat{C}_3^2\phi_1 = (1)\phi_1 + (1)\phi_2 + (1)\phi_3 \\ &= \phi_1 + \phi_2 + \phi_3, \end{aligned} \tag{2}$$

$$\hat{\mathcal{P}}^{E'}\phi_1 \approx (1)\hat{E}\phi_1 + (\varepsilon)\hat{C}_3\phi_1 + (\varepsilon^*)\hat{C}_3^2\phi_1 = \phi_1 + \varepsilon\phi_2 + \varepsilon^*\phi_3, \tag{3}$$

$$\hat{\mathcal{P}}^{E''}\phi_1 \approx (1)\hat{E}\phi_1 + (\varepsilon^*)\hat{C}_3\phi_1 + (\varepsilon)\hat{C}_3^2\phi_1 = \phi_1 + \varepsilon^*\phi_2 + \varepsilon\phi_3. \quad (4)$$

The A SALC in Eq. (2) is real. The two E SALCs [Eq. (3) and Eq. (4)], however, are complex functions and do not coincide with the real UHF orbitals. These complex SALCs agree with BS-UHF orbitals obtained in the case of an applied magnetic field. On the other hand, a set of two real and orthogonal SALCs that spans the E representation can be derived fron Eq. (3) and Eq. (4) by simply adding and substracting the two complex ones. This procedure recovers immediately the real UHF orbitals displayed in Fig. 4.

3 Restoration of circular symmetry: Group structure and sequences of magic angular momenta

In the previous section, we discussed how the BS-UHF determinants and orbitals describe indeed 2D molecular stuctures (semi-classical Wigner molecules) in close analogy with the case of natural 3D molecules. However, the study of the WMs at the UHF level restricts their description to the *intrinsic* (nonrotating) frame of reference. Motivated by the case of natural atoms, one can take a subsequent step and address the properties of *collectively* rotating QWMs in the laboratory frame of reference. As is well known, for natural atoms, this step is achieved by writing the total wave function of the molecule as the product of the electronic and ionic partial wave functions. In the case of the purely electronic or bosonic WMs, however, such a product wave function requires the assumption of complete decoupling between intrinsic and collective degrees of freedom, an assumption that might be justifiable in limiting cases only.

Using the BS UHF solutions, this subsequent step can be addressed by using the post-Hartree-Fock method of *restoration of broken symmetries* [2, 8] via projection (PRJ) techniques.

In this section, we will use the PRJ approach to illustrate how certain universal properties of the CI (exact) solutions, i.e., the appearance of magic angular momenta in the exact rotational spectra, [2, 9, 10, 11, 12, 13] relate to the symmetry broken UHF solutions. Indeed, we will demonstrate that the magic angular momenta are a direct consequence of the symmetry breaking at the UHF level and that they are determined fully by the molecular symmetries of the UHF determinant.

As an illustrative example, we have chosen the relatively simple, but non trivial case, of $N = 3$ electrons. For $B = 0$, both the $S_z = 1/2$ and $S_z = 3/2$ polarizations can be considered. We start with the $S_z = 1/2$ polarization, whose BS UHF solution (let's denote it by $|\downarrow\uparrow\uparrow\rangle$) exhibits a breaking of the total spin symmetry in addition to the rotational symmetry. We first proceed with the restoration of the total spin by noticing that $|\downarrow\uparrow\uparrow\rangle$ has a point-group symmetry lower than the C_{3v} symmetry of an equilateral triangle. The C_{3v} symmetry, however, can be readily restored by applying the projection operator in Eq. (1) to $|\downarrow\uparrow\uparrow\rangle$ and by using the character table of the cyclic C_3 group (see Table 1). Then for the

 ISBN: 978-84-608-6082-2

intrinsic part of the many-body wave function, one finds two different three-determinantal combinations, namely

$$\Phi^{E'}_{intr}(\gamma_0) = |\downarrow\uparrow\uparrow\rangle + e^{2\pi i/3}|\uparrow\downarrow\uparrow\rangle + e^{-2\pi i/3}|\uparrow\uparrow\downarrow\rangle, \tag{5}$$

and

$$\Phi^{E''}_{intr}(\gamma_0) = |\downarrow\uparrow\uparrow\rangle + e^{-2\pi i/3}|\uparrow\downarrow\uparrow\rangle + e^{2\pi i/3}|\uparrow\uparrow\downarrow\rangle, \tag{6}$$

where $\gamma_0 = 0$ denotes the azimuthal angle of the vertex associated with the original spin-down orbital in $|\downarrow\uparrow\uparrow\rangle$. We note that the intrinsic wave functions $\Phi^{E'}_{intr}$ and $\Phi^{E''}_{intr}$ are eigenstates of the square of the total spin operator $\hat{\mathbf{S}}^2$ ($\hat{\mathbf{S}} = \sum_{i=1}^{3} \hat{\mathbf{s}}_i$) with quantum number $s(s+1) = 3/4$, ($s = 1/2$)). This can be verified directly by applying $\hat{\mathbf{S}}^2$ to them.

To restore the circular symmetry in the case of a $(0, N)$ ring arrangement, one applies the projection operator [2, 8, 14],

$$2\pi\mathcal{P}_I \equiv \int_0^{2\pi} d\gamma \exp[-i\gamma(\hat{L} - I)] \ , \tag{7}$$

where $\hat{L} = \sum_{j=1}^{N} \hat{l}_j$ is the operator for the total angular momentum. Notice that the operator $\mathcal{P}_I$ is a direct generalization of the projection operator in Eq. (1) to the case of the continuous cyclic group C_∞ [the phases $\exp(i\gamma I)$ are the characters of C_∞].

The projected wave function, Ψ_{PRJ}, (having both good total spin and angular momentum quantum numbers) is of the form,

$$2\pi\Psi_{PRJ} = \int_0^{2\pi} d\gamma \Phi^{E}_{intr}(\gamma) e^{i\gamma I}, \tag{8}$$

where now the intrinsic wave function [given by Eq. (5) or Eq. (6)] has an arbitrary azimuthal orientation γ. We note that, unlike the phenomenological Eckardt-frame model [13] where only a single product term is involved, the PRJ wave function in Eq. (8) is an average over all azimuthal directions of an infinite set of product terms. These terms are formed by multiplying the UHF intrinsic part $\Phi^{E}_{intr}(\gamma)$ with the external rotational wave function $\exp(i\gamma I)$ (the latter is properly characterized as "external", since it is an eigenfunction of the total angular momentum $\hat{L}$ and depends exclusively on the azimuthal coordinate γ).

The operator $\hat{R}(2\pi/3) \equiv \exp(-i2\pi\hat{L}/3)$ can be applied onto Ψ_{PRJ} in two different ways, namely either on the intrinsic part Φ^{E}_{intr} or the external part $\exp(i\gamma I)$. Using Eq. (5) and the property $\hat{R}(2\pi/3)\Phi^{E'}_{intr} = \exp(-2\pi i/3)\Phi^{E'}_{intr}$, one finds,

$$\hat{R}(2\pi/3)\Psi_{PRJ} = \exp(-2\pi i/3)\Psi_{PRJ}, \tag{9}$$

from the first alternative, and

$$\hat{R}(2\pi/3)\Psi_{PRJ} = \exp(-2\pi I i/3)\Psi_{PRJ}, \tag{10}$$

from the second alternative. Now if $\Psi_{PRJ} \neq 0$, the only way that Eqs. (9) and (10) can be simultaneously true is if the condition $\exp[2\pi(I-1)i/3] = 1$ is fulfilled. This leads to a first sequence of magic angular momenta associated with total spin $s = 1/2$, i.e.,

$$I = 3k + 1,\ k = 0, \pm 1, \pm 2, \pm 3, ... \tag{11}$$

Using Eq. (6) for the intrinsic wave function, and following similar steps, one can derive a second sequence of magic angular momenta associated with good total spin $s = 1/2$, i.e.,

$$I = 3k - 1,\ k = 0, \pm 1, \pm 2, \pm 3, ... \tag{12}$$

In the fully polarized case, the UHF determinant is denoted as $| \uparrow\uparrow\uparrow \rangle$, and it is already an eigenstate of $\hat{\mathbf{S}}^2$ with quantum number $s = 3/2$. Thus only the rotational symmetry needs to be restored, that is, the intrinsic wave function is simply $\Phi^A_{intr}(\gamma_0) = | \uparrow\uparrow\uparrow \rangle$. Since $\hat{R}(2\pi/3)\Phi^A_{intr} = \Phi^A_{intr}$, the condition for the allowed angular momenta is $\exp[-2\pi I i/3] = 1$, which yields the following magic angular momenta,

$$I = 3k,\ k = 0, \pm 1, \pm 2, \pm 3, ... \tag{13}$$

We note that in high magnetic fields only the fully polarized case is relevant and that only angular momenta with $k > 0$ enter in Eq. (13) (see Refs. [15, 16]). In this case, in the thermodynamic limit, the partial sequence with $k = 2q + 1$, $q = 0, 1, 2, 3, ...$ is directly related to the odd filling factors $\nu = 1/(2q + 1)$ of the fractional quantum Hall effect [via the relation $\nu = N(N - 1)/(2I)$]. This suggests that the observed hierarchy of fractional filling factors in the quantum Hall effect may be viewed as a signature originating from the point group symmetries of the intrinsic wave function Φ_{intr}, and thus it is a manifestation of symmetry breaking at the UHF mean-field level.

4 Summary

The analysis presented above concerning the relation between hidden symmetries and emergent signatures (e.g., magic amgular momenta) in the spectra of symmetry-restored mean-field wave functions applies also in the case of configuration-interaction, exact many body wave functions; see the review in Ref. [2]. The CI method requires large-scale, parallel computations, but it has the advantage of providing benchmark results due to the achieved high quantitative accuracy. We refer to this combination of mean-field and CI analysis as computational microscopy [17].

Recently, we have incorporated in our CI computer codes the option of Dirac-delta contact interactions, in addition to the long-range Coulomb one. Thus we have been able to analyze the wave function anatomy of a few untracold fermionic (^{6}Li) atoms in single

 ISBN: 978-84-608-6082-2

and double optical traps, where the formation of quantum Wigner molecules can be associated with the emergence of Heisenberg antiferromagnetic behavior [17, 18, 19] and the creation of highly entangled states; e.g., the celebrated Bell states for two ^{6}Li atoms. (Such behavior was earlier predicted in the case of strongly repelling electrons in double quantum dots [18].) The unprecedented experimental control of the interparticle interaction (from zero to infinite strength) achieved in the case of a few trapped ultracold atoms is enabling investigations of fundamental physics (such as high T_c and 1D and 2D magnetism) from a bottom-up perspective. In addition, in analogy with the field of 2D semiconductor double and triple QDs, it promises technological applications in the area of quantum information and quantum computing. Our computational-microscopy approach can be used to investigate the universal behavior in the strongly correlated regime of both ultracold fermionic or bosonic trapped atoms and confined electrons.

Acknowledgements

This work is supported by the Office of Basic Energy Sciences of the US Department of Energy (Grant No. FG05- 86ER45234).

References

[1] C. Yannouleas and U. Landman, *Spontaneous symmetry breaking in single and molecular quantum dots*, Phys. Rev. Lett. **82** (1999) 5325-5328; (E) Phys. Rev. Lett. **85** (2000) 2220

[2] C. Yannouleas and U. Landman, *Symmetry breaking and quantum correlations in finite systems: Studies of quantum dots and ultracold Bose gases, and related nuclear and chemical methods*, Rep. Prog. Phys. **70** (2007) 2067-2148

[3] I. Romanovsky, C. Yannouleas and U. Landman, *Crystalline Boson phases in harmonic traps: Beyond the Gross-Pitaevskii mean field*, Phys. Rev. Lett. **93** (2004) 230405

[4] I. Romanovsky, C. Yannouleas, L.O. Baksmaty and U. Landman, *Bosonic molecules in rotating traps*, Phys. Rev. Lett. **97** (2006) 090401

[5] C. Yannouleas and U. Landman, *Group theoretical analysis of symmetry breaking in two-dimensional quantum dots*, Phys. Rev. B **68** (2003) 035325

[6] F.A. Cotton, *Chemical Applications of Group Theory*, Wiley, New York, 1990

[7] A.B. Wolbarst, *Symmetry and Quantum Systems*, Van Nostrand Reinold, New York, 1977

[8] P. Ring and P. Schuck, *The Nuclear Many-body Problem*, Springer-Verlag, New York, 1980

[9] S.M. Girvin and T. Jach, *Interacting electrons in two-dimensional Landau levels: Results for small clusters*, Phys. Rev. B **28** (1983) 4506-4509

[10] P.A. Maksym and T. Chakraborty, *Quantum dots in a magnetic-field: Role of electron-electron interactions*, Phys. Rev. Lett. **65** (1990) 108-111

[11] W.Y. Ruan, Y.Y. Liu, C.G. Bao and Z.Q. Zhang, *Origin of magic angular momenta in few-electron quantum dots*, Phys. Rev. B **51** (1995) 7942-7945

[12] T. Seki, Y. Kuramoto and T. Nishino, *Origin of magic angular momentum in a quantum dot under strong magnetic field*, J. Phys. Soc. Jpn. **65** (1996) 3945-3951

[13] P.A. Maksym, *Eckardt frame theory of interacting electrons in quantum dots*, Phys. Rev. B **53** (1996) 10871-10886

[14] C. Yannouleas and U. Landman, *Strongly correlated wave functions for artificial atoms and molecules*, J. Phys.: Condens. Matter **14** (2002) L591-L598

[15] C. Yannouleas and U. Landman, *Trial wave functions with long-range coulomb correlations for two-dimensional N-electron systems in high magnetic fields*, Phys. Rev. B **66** (2002) 115315

[16] C. Yannouleas and U. Landman, *Quantum dots in high-magnetic fields: Rotating-Wigner-molecule versus composite-fermion approach*, Phys. Rev. B **68** (2003) 035326

[17] B.B. Brandt, C. Yannouleas and U. Landman, *Double-well ultracold-fermions computational microscopy: Wave-function anatomy of attractive-pairing and Wigner-molecule entanglement and natural orbitals*, Nano Lett. **15** (2015) 7105

[18] Ying Li, C. Yannouleas and U. Landman, *Artificial quantum-dot helium molecules: Electronic spectra, spin structures and Heisenberg clusters*, Phys. Rev. B **80** (2009) 045326

[19] S. Murmann, F. Deuretzbacher, G. Zürn, J. Bjerlin, S.M. Reimann, L. Santos, T. Lompe and S. Jochim, *Antiferromagnetic Heisenberg spin chain of a few cold atoms in a one-dimensional trap*, Phys. Rev. Lett. **115** (2015) 215301

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

The method of cloud services using for testing in mathematical education

Marina V. Yashina [1], Anastasia S. Dotkulova [2] and Ivan I. Nakonechniy [1]

[1] *Department of Mathematical Cybernetics and IT, Moscow Tech. Univ. of Communications and Informatics (MTUCI) and MADI*

[2] *Department of Mathematics, Moscow Automobile and Road State Tech. Univ. (MADI) and MTUCI*

emails: yash-marina@yandex.ru, asoll05@inbox.ru, razel@bk.ru

Abstract

The paper deals with method of cloud services using for testing process for students specialized of applied mathematics in Moscow automobile and road technical university at mathematical department. The SSSR-education system is developed for preparation of test tasks with Tex compilation at the web-server and using of algorithm of adaptive evaluation of tasks complexity. Also we discuss the advantages of using of cloud testing service structure. This system has been tested on several student groups and shown an increase in student performance at a regular testing.

1. Introduction

Many recent Russian and international conferences are devoted to the development of modern education. Education is a complex socio-technical system, which operates with a large amounts of information, which must be stored and transferred. The modern world, being the world of information technologies, requires a new methods and approaches for the presentation, processing and analysis of information. Using the systems modelling the virtual reality and the Internet-resources, the educational process has taken a step to a new spiral of evolution. The use of computer technology to address applications tasks is a required attribute of the professional activities of any specialist. Therefore, preparation of students in higher educational institutions, especially those who study applied sciences, is not possible without the use of modern education technologies. First of all, it relates to the use of

 ISBN: 978-84-608-6082-2

intelligent knowledge control systems and computer information and communication technologies in the education process. The use of mobile computing devices (tablets, ultrabooks, e-books, smartphones, etc.), permanent broadband access to the Internet which accelerates the creation of cloud computing and the ability to connect to the remote data, now becomes a more affordable everyday reality. Cloud services is one of the most promising directions for the development of modern information technologies. Cloud services is a new service, which implies the remote data processing and storage. Using the "cloud" you can gain access to the information resources of any level and any power, with the limitation of the access rights to the information that is stored on the server, using only the access to the Internet. Cloud services can be divided into the three main categories: infrastructure as a service; platform as a service; software as a service. As an example of the use of cloud computing technology in education we can mention a personal accounts for students and teachers, electronic journals, data storage, where students may exchange and search for information, where students may resolve certain educational tasks even in the absence of a teacher or under his leadership. Our scientific group under the leadership of prof. Buslaev A.P. since the beginning of 2000s is engaged in the development of the systems of distributed monitoring and management of complex social systems (Bugaev, Buslaev, Kozlov, Yashina, 2011). In this work we consider the issues of information systems in the field of education and algorithms for extracting of the useful information from a semi-structured source on the example of our system of educational process maintenance - SSSR-Ed (Buslaev, Burikova, Guseva, Nakonechniy, Yashina, 2013).

2. Methods of knowledge assessment based on test results

2.1. The use of information technologies for knowledge assessment purposes

In recent years the computer technology has become an integral part of the modern teaching methods. The solution to the problem of ensuring the individualization of learning, improving the activity and the quality of education can be achieved only on the basis of the limited use of electronic and computer equipment in the education process together with the traditional methods of education science. Information capabilities and performance of modern computers offer an unlimited scope for pedagogical creativity of teachers, allowing to upgrade the old technologies and offering the new technologies and forms of education. The advantages of the computer systems to control the quality of knowledge is the operativeness and performance of testing data processing. In the education process testing helps to check the learning of this or that material, both by the teacher and by the student. For example, if a student does not know how to find derivatives, it does not make sense for him to start learning integrals. In addition, testing helps at entrance examinations as a first phase: students entering a university, have different knowledge in a given area, so the test-

ing helps to form a preliminary groups by skills, which in the future will be divided into the interest areas. In the educational process itself testing serves as an objective assessment, not dependent on the subjective opinion and relationships between the student and the teacher. During sessional examinations testing helps to separate successful students from the ones falling behind, which is important in large flows. Testing is used to evaluate not only the work of students, but also the effectiveness of teaching methods and the work of the teacher himself. The relevance of testing is also contingent upon the flexibility of the system, different solutions to problems are possible, which in turn helps to assess students from various points of view. However, testing can not be considered the major assessment component of the education. Primarily for applied and arts professions that require not only testing knowledge, but also the skills of a person. Therefore, in order to prepare specialists in the field of applied mathematics was generated a cloud of tests.

2.2. Knowledge assessment models

Currently, there are different models of knowledge assessment, divided into three major classes: models for assessing the level of knowledge, knowledge diagnostics models and recognition models. The models to assess the level of knowledge aimed at obtaining an integrated, quantitative mark of a testee - points. Diagnostics models allow to determine problems in knowledge. Recognition models set the task to allocate the testee after the test to one of the predefined classes, such as "certified" or "not certified". A simple model of knowledge assessment is more often used to assess the knowledge of large flows (Zayceva, Prokofyeava, 2004). Student's response to each task is evaluated based on the two-degree scale (correct or incorrect). The mark is awarded by calculating the value of the parameter R:

$$R = \frac{\sum_{i=1}^{k} R_i}{n}, \tag{1}$$

where R_i is student's correct answer to the task; k is the amount of correct answers of n the presented ones ($k?n$), which, as a rule, is an integer (or it is rounded according to the rules of mathematics). Final mark, as a rule, is determined by the following formula:

$$I = \begin{cases} 1, R \leq c_1 \\ 2, c_1 \leq R \leq c_2, \end{cases} \tag{2}$$

where I is a final mark, $\{c_1, c_2, , c_M\}$ is the vector of boundary values (which was predetermined), M is the highest possible mark. To assess the knowledge of applied mathematicians

 ISBN: 978-84-608-6082-2

modern mental test theory is usually used - Item Response Theory (Rudinskiy, Grusheckiy, 2003). In contrast to the classical theory, where the individual score of the tested student is viewed as a constant number, in IRT it is interpreted as some variable. The initial value of the parameter is obtained directly from the empirical test data (mean of sample) As an example of statistical assessment model let us review the method, considering the type of data distribution. The main idea of this method is that with a sufficiently large test sample, provided to the student (with the number of test questions not less than 50), the students answers distribution functions, characterized by different levels of knowledge, tend to be well-defined in accordance with laws of distribution, for example, a normal distribution. So, if there is full and deep knowledge, the distribution of answers is close to an exponential one with parameter 1, and in the case of the complete lack of knowledge - to exponential distribution with parameter 0. If the student demonstrates not excellent, but smooth and confident knowledge of the entire test topic, the distribution of his answers will be close to normal with a distinct maximum and a relatively small dispersion, whereas if there are significant gaps on a specific topics, this distribution will be close to normal with slightly marked peak and a large dispersion. At the same time, when a student tries to guess the correct answers (i.e., in case of a random selection) the distribution will be close to equal one. The basic idea of the proposed algorithm is the following. When conducting automated testing of knowledge, taking into account the answer options selected by the students, obtained answers are grouped and constructed a frequency polygon of their distribution. Null and alternative hypothesis of an exponential, normal and equilibrium distribution of the sample of answers are consistently nominated. The nominated hypothesis are checked with the appropriate consent criteria; the hypothesis, which has the highest degree of consent, is selected for further analysis. Taking into account the parameters of the accepted law of distribution, the size of the sample and the required probability belief, the size of the confidence interval is calculated, which is projected onto the reference assessment scale for the selection of the evaluation. In the case if the confidence interval is completely placed in the area between the two neighboring marks, the higher one is awarded. The situation in which the confidence interval overlaps the two neighboring marks suggests on the lack of certainty of the test results. This uncertainty can be resolved either by providing the trainee an additional test questions with subsequent repetition of the calculation with the increased sample size, or by awarding a lower mark, corresponding to the overlap border of the areas. In our work we used the algorithm of adaptive knowledge testing (Bessarabov, Bondarenko, Kondratenko, Timofeev, 2016). This algorithm consists the following steps: Organization of a homogeneous group of students for testing; Generation of organizational test parameters; Taking into account the mean square error of preparedness, stochastic approximation is processed and optimal value of complexity of the test tasks is calculated; The generated test is presented to the test group; After the test, statistical processing is produced of obtained matrix of responses and value of the complexity of the test tasks is re-

calculated; Adaptation parameters for a test group of students are computed to the test for the next iteration; the difference between the resolution and the ability to test a calculation error in the results is calculated; 8) The total score for each tested student is calculated. Statistical methods are also applied for the assessment of the tests quality, reliability of the testing, forecasting of the results of tests.

3. Cloud testing method based on SSSR-Ed

3.1. SSSR-Ed project

In these days cloud technologies are increasingly introduced in all spheres of life. Thanks to such a rapid development the educational process is becoming more flexible and affordable. In this article it was decided to tell about the cloud technology used in the education of applied mathematicians, namely, to consider some of the modules of the SSSR-Ed system. As a part of the REC project "Theoretical and applied issues of creating systems of intelligent monitoring and control of distributed processes" under the leadership of Buslaev A.P. was developed a client-server system SSSR (1). SSSR system (smartphone, server, student, routing) is a technology of automated data collection using mobile devices, smartphones and tablets, with the sending of multimedia information to the server in a structured way. Were developed the theoretical bases and client-server systems technologies for distributed monitoring, and later the SSSR-Ed system was developed for the monitoring of educational process in the University. Main tasks of the SSSR-EJ are the monitoring of the education process in real time, its synchronization with the schedule, introduction of statistics on the attendance and performance of students. The system provides an ability to record the situation at the classes in a real-time mode, even if the University is located in multiple buildings, because the data transfer is carried through the 3G connection. EJ-SSSR database collects information on attendance of each training session scheduled by the academic office, on the type of this session (lectures, practical classes, laboratory work), on the multimedia of this session, marks received by the students. After connecting to the server, in an online mode the smartphone receives the real-time information about the session. Therefore, you can not enter the relevant data neither before nor after the education session, allowing their full credibility and hampering the implementation process. The main advantage over the similar systems is the lack of binding to a specific operator workstation in order to implement interaction between staff and the base. It was decided to use a variety of mobile devices running Android and iOS operating systems as client devices. University staff have the ability to transmit data through a smartphone using the GPRS, EDGE, 3G and Wi-Fi IEEE 802.11 b/g technologies. Thus, multimedia data can be transmitted from any point of the University, even if the University is geographically spread, and various buildings are located at a considerable distance from each other. Client device requirements are very moderate, allowing to use budget-friendly smartphones and tablets. Android devices need

to have the operation system version 2.1 and above. For iOS devices - version 4 and above. The disadvantages of the system are the need for a permanent Internet connection (if there is no connection to the Internet it freezes and is unable to update the data) and the lack of local mode.

Figure 1: Logical scheme of interaction.

Currently this system is tested on the basis of the MKiIT (Mathematical Cybernetics and Information Technologies) academic department of the Moscow Technical University of Communications and the Advanced Mathematics academic department of the Moscow Automobile and Road Construction University. In addition, we have filed an application for a patent for a utility model on the topic: "The method of automated maintenance of the educational process". SSSR-Ed system includes several components: client part, server part, cloud services. The client part of the system is installed on the Android operating system and is necessary to transfer the information in real time. The Server part is a hosting, which stores all the information transmitted from the client device, and is necessary for

the output of the obtained information. Cloud services allow you to collect and submit additional materials necessary for the successful studies.

Figure 2: Conceptual scheme of interaction processes in the SSSR-Ed project.

3.2. Method of test generation and database structure

In the Moscow Automobile and Road Construction University the department of "Advanced Mathematics" for more than 20 years is engaged in the research of automated systems and information processing tools, it models and develops the mathematical software for the new generation of computers, it develops software and information support for computer networks, automated systems of computing complexes, services, operating systems and distributed databases. The specialized system of typesetting TeX, which specializes in the input of mathematical functions, and has an impressive list of symbols and mathematical operands is also widely used. The TeX system is widely used at the department of "Higher mathematics" for writing of tutorial materials, manuals and assignments for students. TeX convenient system for preparation the layout of the text containing mathematical formulas.

```
\documentclass{article}¶
\usepackage[cp1251]{inputenc}¶
\usepackage[russian]{babel}¶
\usepackage{geometry}¶
\usepackage{amsmath}¶
\begin{document}¶
\vskip 5.5cm   \hspace{2.5cm}¶
{\vskip 0.5cm }\centerline{\bf\hskip1.5cm  DISCRETE MATH}¶
{\vskip0.0cm }\centerline{\bf\hskip1.5cm INDEPENDENT WORK. VARIANT №1}¶
\vskip 0.5cm¶
\zadan{ 1}{¶
Change the order of integration: $\int_0^1 dx\int_{1-x}^{\sqrt{1-x^2}}f(x,y)dy$.\\¶
{\bf1.}$\scriptstyle \int_0^1 dy\int_0^1 f(x,y)dx$¶
{\bf2.}$\scriptstyle \int_0^1 dy\int_0^{1-y} f(x,y)dx$¶
{\bf3.}$\scriptstyle \int_0^1 dy\int_{\sqrt{1-y^2}}^{1-y}f(x,y)dx$¶
{\bf4.}$\scriptstyle \int_0^1 dy\int_{1-y}^{\sqrt{1-y^2}}f(x,y) dx$¶
{\bf5.}n1-n4 are false¶
}\\¶
\zadan{ 2}{¶
Whether true identity. $(A \cup B) \times C = (A \times C) \cup (B \times C)$. The answer to
justify. \\¶
{\bf1.} Yes¶
{\bf2.} No¶
{\bf3.} I don't know}\\¶
\zadan{ 3}{Whether couples are isomorphic representations of } ¶
\begin{figure}[h]¶
\center{\includegraphics[width=1\linewidth]{image}}¶
\label{ris:image}¶
\end{figure}\\¶
\end{document}
```

Figure 3: Listing of testing scheme in Latex.

DISCRETE MATH
INDEPENDENT WORK. VARIANT №1

1. Change the order of integration: $\int_0^1 dx \int_{1-x}^{\sqrt{1-x^2}} f(x,y)dy$.

1. $\int_0^1 dy \int_0^1 f(x,y)dx$ **2.** $\int_0^1 dy \int_0^{1-y} f(x,y)dx$ **3.** $\int_0^1 dy \int_{\sqrt{1-y^2}}^{1-y} f(x,y)dx$ **4.** $\int_0^1 dy \int_{1-y}^{\sqrt{1-y^2}} f(x,y)dx$ **5.** n1-n4 are false

2. Whether true identity. $(A \cup B) \times C = (A \times C) \cup (B \times C)$. The answer to justify

1. Yes **2.** No **3.** I do not know

3. Whether couples are isomorphic representations of

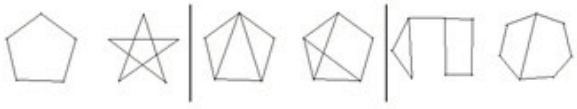


Figure 4: Testing scheme after compilation.

 ISBN: 978-84-608-6082-2

TeX system is a specialized programming language. Donald Knuth not only came up with the language itself, but also wrote a compiler for it that works smoothly on different computer configurations. This system serves as a basis for other publishing systems used in practice. More precisely, every publishing system based on TeX is a package of macros (macro package) of this language. Let us underline the main advantages and disadvantages of this product in comparison with similar ones. Advantages of this language include:

- It is free and open;
- More features compared to the classic word processors. For example, auto-compiled contents, bibliographies, the possibility of merging documents, comfortable numbering of formulas, figures, simple syntax of mathematical formulas, cross-references, etc.;
- Well-structured text;
- Wide variety of styles to prepare different documents;
- 100% backwards compatibility between versions;
- Convenient tags for the formation of the text;
- The possibility of collective work;
- The system to monitor versions of the document;
- LaTeX is actually designed specifically for the physical-mathematical articles that contain many formulas;
- Portability of the document on different platforms.

The following should be considered as the weaknesses:

- The complexity of the mutual integration with documents created in ex and MS Word;
- The need to study the syntax of the layout that makes it difficult to work on the initial stages of education.

We developed the online TeX compiler for our SSSR-ED system to make work with mathematical formulas faster (3, 4). Students can access all LaTeX power and functions without installing TeX-software on their gadget, usually their smartphones or tablets. Device should have internet-access. Students may use the compiler any time, except the time of exam or the planned test work. Any student can view the computing log file or share the results with the fellow students. The Online TeX-system recommended itself as very powerful and useful tool among the teachers. The bank of test tasks on all the main topics of the mathematics course for the occupations of III generation was created in order to monitor

performance and learning achievements. The bank includes more than 3,000 tasks stored in .lat format that may be opened in any text notepad. The "Test Maker" app was developed for convenience and correct configuration of tasks for teachers, preparing assignments for the exam to a large audience. The teacher, on the basis of methodological considerations, chooses topics and types of tasks from the proposed menu, and the software application generates random options, differing by the numeric constants, and generates a ready-made forms of examination tasks that have been formatted in LaTex (5).

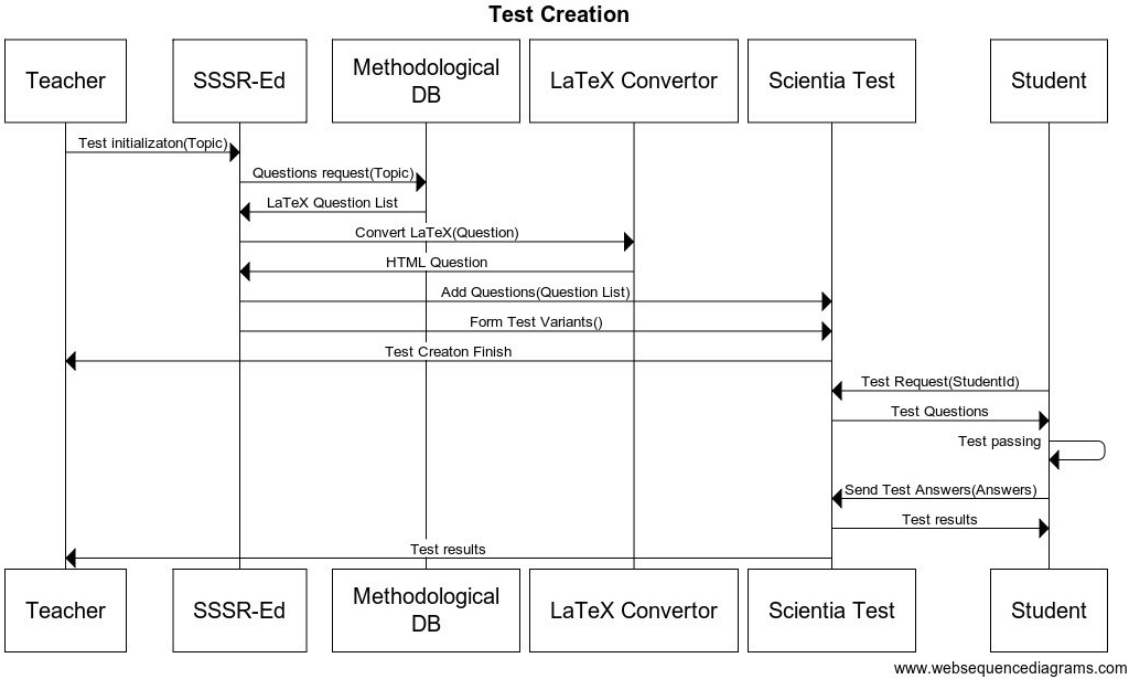


Figure 5: Preparation of assignments.

Later the generation of texts was expanded and implemented the Scientia-test testing system. This multifunction system provides support for monitoring students' progress at any one time in case if the students have the access to the system. Access is granted through the username and password individually generated for each student. With such identification any student can view his results and pass a sample test at any desired time and from any device.

3.3. Scientia-Test web-service

Scientia-Test is a universal tool for the automation of the intermediate control of students' knowledge. It gives teachers an opportunity to fill the knowledge database, and students have an opportunity to take the test on-line. This system is used in MADI only to solve text tasks, but it can be used for a wide range of tasks:

- inspection and control of knowledge of university students at the intermediate control of knowledge;
- self-control of the students (there is a "mock" test in the systems, which allows any student to prepare for the test at home);
- teachers can create test materials (both using the web-interface and through input of the introduction into the designated form);
- creation of a single test base for a wide range of University subjects;
- preparation to the state accreditation from the University.

The program (website) to create scientia-test tests has proven its effectiveness by ensuring:

- effective automation of carrying tests and evaluation of the test results through a wide range of features;
- convenient operation thanks to the modern user interface;
- elimination of physical and time-related costs due to the automated processing of the results;
- independent creation of tests by teachers and, if necessary, independent work of the students with the program (in case of self-control).

 ISBN: 978-84-608-6082-2

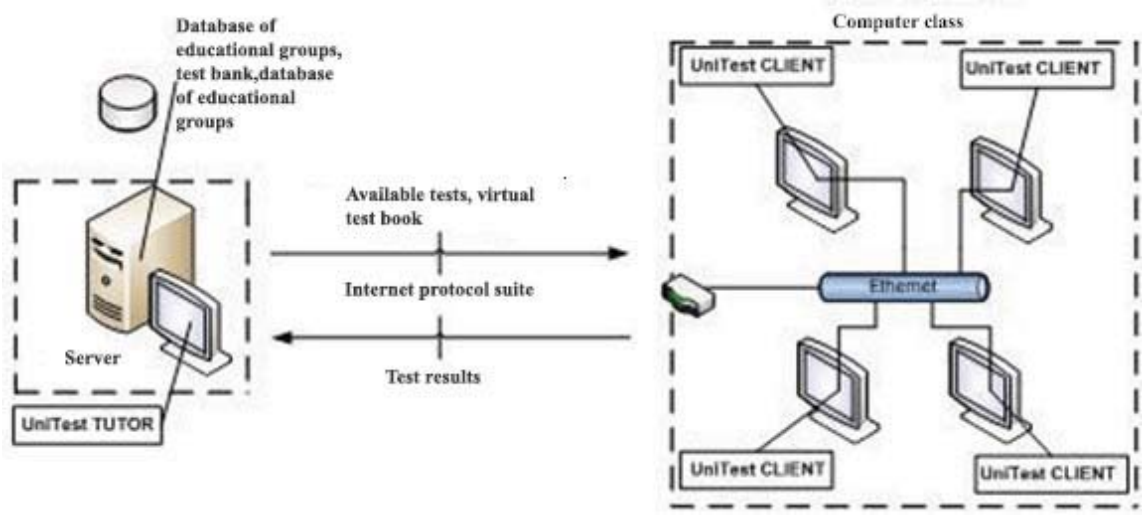


Figure 6: Testing scheme.

We have made an add-on for the system that allows uploading tasks to the testing database automatically, on the basis of files prepared in the latex format (6). This add-in also allows the use of mathematical formulas in tests. With the module that simulates user activities, assignment options are added to the testing system, which are then automatically generated and compiled into a test. Thanks to the Internet now there is an opportunity to carry tests not only in selected groups, but also in large streams in the form of examinations and assignments. Test results are generated immediately and then are sent to the cloud, where they are stored in an electronic attendance-based journal. Attendance card contains statistics of marks both in the whole group and for a single student. The results are read from the attendance cards, and the statistics is generated (7).

3.4. Cloud structure of the SSSR service

In the course of operation it was found that the SSSR-Ed, Scientia-Test systems and the Tutorial database mutually complement each other in the work of maintaining the educational process. The problem was in their isolation and lack of the tools for mutual synchronization between them. SSSR-Ed had an open platform, which was a web service, and it was used as a basis to construct this cloud solution.

MADI

CARD

ATTENDANCE AND PERFORMANCE OF EDUCATIONAL TASKS

St. gr 16AC1 Term 1 2014-2015 ac.y. Teacher Buslaev Aleksandr Pavlovich

Subject Mathematics 1

№ n/n	FULL NAME	The percentage of attendance	Attendance and actual performance																												
			04.09(Lec)	04.09(prac)	08.09(prac)	11.09(prac)	18.09(Lec)	18.09(Lec)	22.09(prac)	25.09(Lec)	25.09(Lec)	02.10(prac)	02.10(prac)	06.10(Lec)	09.10(Lec)	09.10(prac)	16.10(Lec)	16.10(Lec)	20.10(Lec)													
1	Aduev Rustam Ruslanovich	82,35%			40%				50%					N	N	55%			N													
2	Andreev Gleb Sergeyevich	88,24%			40%				49%					N		51%			N													
3	Asmakovsky Vladislav Sergeevich	88,24%			63%				65%					N		70%			N													
4	Barinov Alexander Romanovich	52,94%			20%		N	N	30%	N				N	N	33%	N	N	N													
5	Voronov Petr Olegovych	88,24%			35%				36%					N		42%			N													
6	Zaripov Bulat Marselevich	88,24%			45%				51%					N		52%			N													
7	Kazankov Andrei Dmitrievich	64,71%	N	N	51%				56%			N	N	N		58%			N													
8	Kashtanov Kirill Viktorovich	58,82%			N	N			48%			N	N	N	N	40%			N													
9	Korostelev Demid Alexandrovich	47,06%			69%	N	N	N	74%	N		N	N	N	N	65%			N													
10	Lebedev Andrey Pavlovich	52,94%			53%		N	N	57%	N		N	N	N	N	69%			N													
11	Lysenko Egor Dmitrievich	88,24%			48%				52%					N		54%			N													
12	Myshkovets Alexander Mihaivovich	88,24%			61%				70%					N		75%			N													
13	Pavlenko Ilia Vladimirovich	88,24%			52%				63%					N		81%			N													
14	Pereguda Ilia Stanislavovich	17,65%	N	N	N	N	N	N	44%	N	N	N	N	N		38%	N	N	N													
15	Puryavkin Dmitry Mikhailovich	76,47%			55%				56%			N	N	N		62%			N													
16	Safiullin Ildar Damirovich	82,35%			42%				40%	N				N		59%			N													
17	Sigankov Nikita Dmitrievich	82,35%			60%			N	72%					N		85%			N													
18	Soshin Eugene Romanovich	41,18%			46%		N		56%	N	N	N	N	N	N	63%	N	N	N													
19	Tolstov Evgeniy Andreevich	82,35%			61%			N	60%					N		69%			N													
20	Tsaplin Konstantin Olegovich	70,59%			49%			N	52%	N	N			N		56%			N													
	Attended		18	18	18	17	15	13	20	13	17	13	13	0	14	20	17	17	0												

Figure 7: Attendance card in cloud storage.

3.4.1. Integration of the Tutorial Database into cloud

Tutorial database represents the system containing of the two modules. The database itself, containing the list of questions in the TeX format, distributed by topics, and the utility yo generate unique tests. As an important part of the integration it was decided to abandon the utility of test generation and replace it with a centralized system - Scientia-Test. We started using the SSSR-Ed service for their mutual integration. The interaction process is detailed below:

1) The student addresses a request to the SSSR-Ed system to provide the test;
2) SSSR-Ed extracts data from the Tutorial DB;
3) SSSR-Ed transfers the obtained data to Scientia-Test;
4) Scientia-Test generates a test for the student;
5) SSSR-Ed uploads the test results to its DB.

In the course of integration it was found out that the Scientia-Test system has a number

of drawbacks, in particular, this system does not support any one of the languages for the visual representation of mathematical formulas.

3.4.2. The interaction of SSSR and Scientia-Test

In the course of operation there was a problem: Scientia-Test system does not support any one of the formulas visual representation languages, the tool to visually represent formulas (TeX) was integrated into the SSSR-Ed system. The developed system retrieves the formula text from the Tutorial database, processes it in LaTeX and generates an image that is to be transferred to the Scientia-Test system. Because the Scientia-Test does not contain an open service or the tools to import or export data, it was decided to use the Selenium suite to upload and download data from this system. Selenium is a tool for automated browser control. Automation of web-application testing is the most popular selenium area of application. However, using selenium it is possible to automate any other routine actions carried out through the browser. Using this package of tools we managed to simulate the actions of the Scientia-Test database administrator. When loading the test data, the SSSR-Ed system breaks the texts of the questions from the Tutorial database into components. The texts of the assignments are transferred directly into the form, and texts, containing formulas, are first sent to the module generating images, after which the link to such image is transferred to the Scientia-Test.

3.4.3. Fraudulent protection

SSSR-Ed system supports the protection and access control of the users in the process of testing. SSSR-Ed database contains data on the student, including his phone number. For the implementation of the central testing we created a dedicated module, which allows to allocate each student a unique temporary key and generate a unique test, as well as checks the device from which the testing is performed (whether it is a computer in the classroom, in which this test is carries, whether this authorization key was used from another device). The data of the previous tests in the same way are stored in the SSSR-Ed DB, which allows you to avoid issuing the same question to the same student in the case of a test-out exam. Interaction process is as follows:
1) The teacher launched centralized testing;
2) SSSR-Ed retrieves questions from the tutorial database;
3) SSSR-Ed converts TeX into the format suitable for Scientia-Test;
4) Scientia-Test generates tests;
5) SSSR-Ed sends keys to the mobile phones of the students;
6) Students gain access keys (access keys are active only during the specific amount of time allocated for the testing);
7) Students perform the tasks;

8) The results are uploaded to the SSSR-Ed database.

4. Analysis of the implementation of this testing system in the educational process

During the 2014/2015 academic year based on the AM academic department in the MADI and MCIT academic department in the MTUCI was implemented a regular testing for the disciplines: Discrete mathematics, Structures and algorithms of data processing, and FLP. Recognition models place the task to allocate the testee after the testing to one of the predefined classes, for example, after testing to one of the predefined classes, for example, "certified" or "not certified" (8).

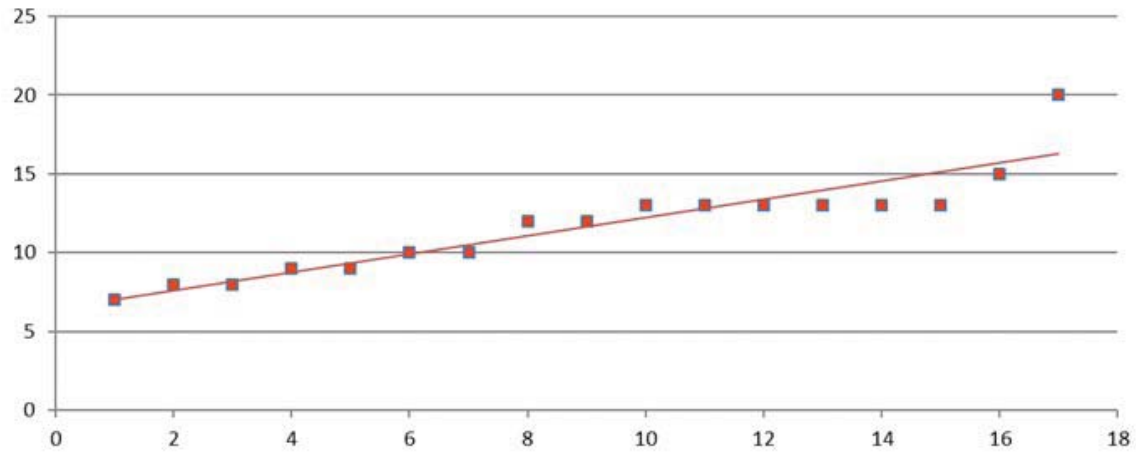


Figure 8: Evaluation criterion: "certified" and "not certified".

The initial value of the parameter is obtained directly from the empirical test data (mean of sample) (9). Next, the statistical models are applied to the obtained results.

Evaluated 10 student groups consisting of 20-25 people each. Among five groups the testing was conducted once a month regularly during the two semesters. The other five groups learned in a regular mode and the testing was conducted only at the end of the school year.

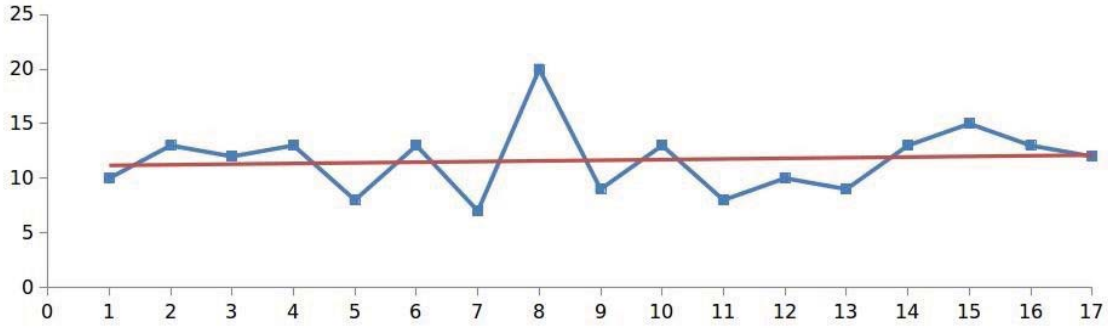


Figure 9: Schedule of assessment IRT model.

The testing lasted for one hour and included 20 questions. The response to each question was evaluated based on the 5-mark system. In the end all the test results were brought to 100% (points) system of calculation. Marks for testing are calculated by the formula (1). The testing helped groups to carry out separate work, which helped to quickly determine whether the material was learned and understood or not. The testing results themselves were analysed by the several models. At first, the simple model for knowledge assessment was considered. A point chart with the number of correctly resolved questions is constructed based on the data obtained by the model (2) (8). The approximating straight line is constructed on the chart, created on the basis of the current average score of students in relation to the number of tasks. The more complex IRT model is used as the second step in the knowledge evaluation (9). Testing marks are calculated by the formula (2) and are not added to all the other ones. In each test the student can get from 0% to 100%, therefore, receiving an assessment mark from "2" to "5". The scale knowledge evaluation is the following: 1) for each task in the test the student gets 5% (points):
- 1 point - provided the answer without any clarifications;
- 2 points - incorrect answer, but the way of finding solution is correct;
- 3 points - provided only half of the solution;
- 4 points - a small accuracy mistake in the solution;
- 5 points - correct and detailed answer.
2) by the 100-point system the marks are awarded by the following criteria:
0-50% - mark 2
51-70% - mark 3
71-85% - mark 4
86-100% - mark 5.

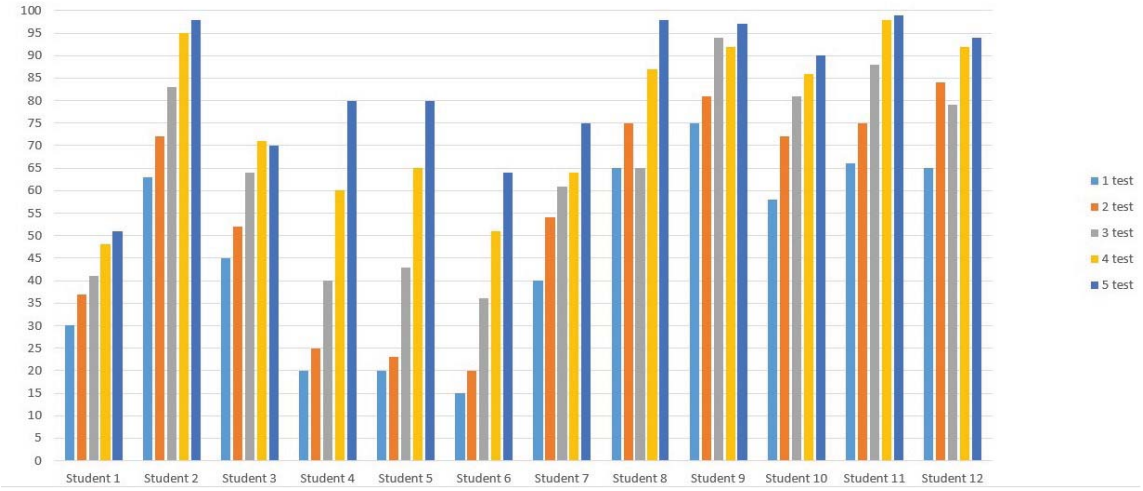


Figure 10: Dynamics of the results of the regular testing of students.

Three students were selected from each of the groups, in which the tests were conducted, to compare the results dynamics of the student compared to himself.

Above there is a statistics on the dynamics of the completion of testing by the individual students from the group, in which the testing was conducted on a regular basis (data is given in percentage) (10). Students of the "average" and "below average" level were chosen deliberately, because the dynamics is less brightly expressed for the more successful students. As can be seen from the diagram, in case of testing on a regular basis the performance indicator either remains unchanged in an average or expresses a positive dynamics.

By the results of the monitoring of student behavior all students can be divided into 4 groups:

- Group A: did not pass any tests during the semesters;
- Group B: the group of students, in which there was no regular testing and the students did not attend the classes;
- Group C: students passed tests not on a scheduled basis (group of students, who often missed classes);
- Group D: group of students, who went though testing on a regular basis.

Using such a separation the average score of the final testing was displayed (11). As can be seen from the graph, the students who has testing on a regular basis, on average have a higher score than the others.

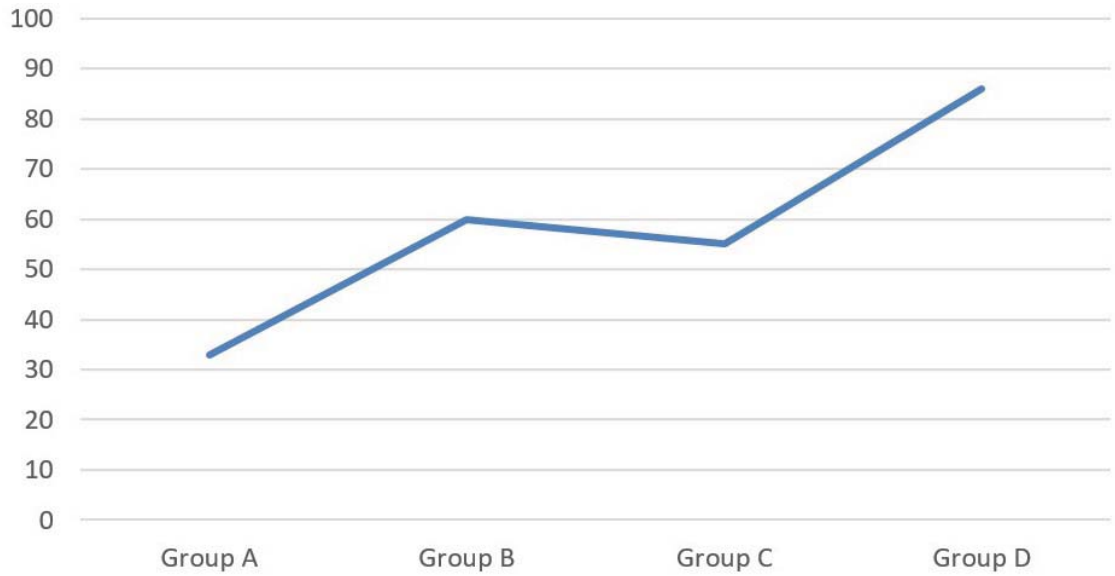


Figure 11: Average final testing of students.

Conclusion

The significant optimization of centralized testing procedures has been performed. Online TeX compiler allowed to normalize of tasks data base. The SSSR-Ed system gives possibilities to students to use their mobile devices as a testing tool. Adaptive knowledge testing increased the adequacy and uniformity of distribution among students by calculating of relative magnitude of the task complexity. The use of cloud technologies is introduced into the educational process, including the education of applied mathematicians, with a delay and yet fails to find a wide application. Cloud technologies provide students with the opportunity to interact and to work collaboratively with the continuously expanding social circle regardless of their location. These technologies deliver education materials through the most cost-effective and reliable way, standing out by the ease of distribution and updating. Cloud computing technologies represent a new way to organize the education process and offers an alternative to the traditional methods of education, creates opportunity for personal education, team teaching and interactive classes. The main advantages of the use of cloud computing technologies in education are not only the reduction of costs for the purchase of necessary software, the more efficient and high quality educational process, but also the fact that students will be more prepared to the life in the modern information society.

References

[1] L.V. Zayceva, N.O. Prokofyeava, *Models and methods for adaptive control of knowledge.* Educational Technology & Society. **7(4)** (2004).

[2] E.D. Rudinskiy, S.V. Grusheckiy *Statistical methods for derivation of automated testing results.* Information Technology in Education, 2003 International Conference (2003).

[3] C. Romero, S. Ventura, M. Pechenizkiy, R. S. Baker *). Handbook of educational data mining*. CRC Press.

[4] A. S. Bugaev, A. P. Buslaev, V. V. Kozlov,M. V. Yashina *Distributed problems of monitoring and modern approaches to traffic modeling*. Intelligent Transportation Systems (ITSC), (2011) 14th International IEEE Conference on (pp. 477-481). IEEE.

[5] N.A. Bessarabov, A.V. Bondarenko, T.N. Kondratenko, D.S. Timofeev *Algorithmic support of adaptive testing knowledge system* Journal "Software & Systems", 1(113), (2016) 68–73.

[6] A.P. Buslaev, R.G. Abyshov, U.D. Kupriyanov, M.V. Yashina *A distributed system for monitoring road maintenance* Vestnik MADI 1(24) (2011) 79 - 85.

[7] A.P. Buslaev, T.A. Burikova, A.S. Guseva, I.I. Nakonechnyy, M.V. Yashina *Application of information - computer networks for monitoring of complex socio-technical processes for the example of evaluating the quality of maintenance of knowledge in mathematics in high school. Part 2. Technology certification.* Handbook, Tehpoligraftsentr (2013) 72.

Volume V

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Dynamical Study of Gursey Instantons with Bichromatic Force

Fatma Aydogmus[1] and Eren Tosyali [2]

[1] *Department of Physics, Istanbul University, Istanbul, TURKEY*
[2] *Vocational School of Health Services, Istanbul Bilgi University, Istanbul, TURKEY*

emails:
fatma.aydogmus@gmail.com, fatmaa@istanbul.edu.tr,
eren.tosyali@bilgi.edu.tr

Abstract

Instantons have finite action with zero energy. They have been considered as configurations of quantum fields providing a tunnelling effect between the vacuums which have different topologies in space-time. In this study, we investigate the transition of four dimensional Gursey instantons from regularity to chaos under the bichromatic potential.

Key words: instanton, spinor, nonlinear dynamic, chaos.
MSC2000: AMS Codes (optional)

1. Introduction

Solitons were discovered in the 19th century as undissipated surface waves on water and realized to obey nonlinear wave equations [1]. Since the discovery of soliton much progress has been made on the subject of nonlinear phenomena especially soliton dynamics. During the past forty years a rather complete description of solitons has been developed by the productive collaboration of mathematicians and physicists. In mathematical physics the amount of information on nonlinear wave phenomena obtained using solitons is quite high. Today it is known that solitons play an important role not only in nonlinear physics and mathematics but also in fiber lasers and communication engineering. There are four leading solitonic characters: instanton, monopole, vortex, and kink ones. Instantons are classical topological solutions with zero energy and finite action of the Euclidean version of field equations of any given model [2-4].

 ISBN: 978-84-608-6082-2

Very recently the dynamical nature of dissipative Gursey Model has been studied to check the stability of Gursey instantons against external forcing and damping and chaotic instantons observed depending on system parameter values [5]. In this paper, we investigate the regular and chaotic solutions of Gursey Model under the bichromatic potential to get more information about the quantum dynamics of spinor type instantons in vacuum.

2. Model

The Gursey wave equation [6] is described by the conformal invariant Lagrangian

$$L = i\bar{\psi}\partial\psi + g(\bar{\psi}\psi)^{\frac{4}{3}}, \tag{1}$$

where the fermion field ψ has scale dimension $\frac{3}{2}$ and g is the positive dimensionless coupling constant. The conformal invariant spinor wave equation that follows the above Lagrangian is

$$i\partial\psi + g(\bar{\psi}\psi)^{1/3}\psi = 0. \tag{2}$$

In Ref. [7], for the spinor type instanton solutions, it has also been shown that $\bar{\psi}\psi$ is related to spontaneous symmetry breaking of the full conformal group and is characterized by being invariant under the transformations of a special subgroup [8], which in turn reflects the final symmetry properties of the ground state of the system. That is

$$R_\mu(\bar{\psi}\psi) \equiv \frac{i}{a}\left[\frac{a^2 - x^2}{2}\partial_\mu + (x.\partial + 2d)x_\mu\right](\bar{\psi}\psi) = 0. \tag{3}$$

Following the proposal of Ref [8] an operator $\left(R_\mu\right)$ was introduced

$$R_\mu = \frac{1}{2}\left(aP_\mu + \frac{1}{a}D_\mu\right), \tag{4}$$

depending momentum (P_μ) and conformal scale invariant (D_μ) operators in the four dimensional Euclidean space-time. Here a is a parameter with the dimension

of length. Then, one finds $\bar{\psi}\psi = \dfrac{a}{g\left(a^2 + x^2\right)}$ as a solution. This solution is related to the special case (instanton) [7] of Euclidian configuration of Heisenberg ansatz [9]

$$\psi = \left[ix_\mu \gamma_\mu \chi(s) + \varphi(s) \right] C, \tag{5}$$

where C is an arbitrary spinor constant; $\chi(s)$ and $\varphi(s)$ are real functions of $s = x_\mu^2 = r^2 + t^2 (x_1 = x, x_2 = y, x_3 = z, x_4 = t)$ in the Euclidean space-time, *i.e.* $r^2 = {x_1}^2 + {x_2}^2 + {x_3}^2$. Inserting Eq. (5) into Eq. (2), with

$$i\partial\!\!\!/\psi = i\gamma_\mu \partial_\mu \psi = \left[-4\chi(s) - 2s\frac{\mathrm{d}\chi(s)}{\mathrm{d}s} + 2ix_\mu \gamma_\mu \frac{\mathrm{d}\varphi(s)}{\mathrm{d}s} \right] \bar{C}C \tag{6}$$

and

$$g(\bar{\psi}\psi)^{1/3}\psi = \left[igx_\mu \gamma_\mu \chi(s)\left(s\chi^2(s) + \varphi^2(s)\right)^{\frac{1}{3}} + g\varphi(s)\left(s\chi^2(s) + \varphi^2(s)\right)^{\frac{1}{3}} \right] (\bar{C}C)^{1/3} C, \tag{7}$$

where

$$(\bar{\psi}\psi)^{1/3} = \left(s\chi^2(s) + \varphi^2(s)\right)(\bar{C}C)^{1/3}, \tag{8}$$

one obtains the following nonlinear differential equations system

$$4\chi(s) + 2s\frac{\mathrm{d}\chi(s)}{\mathrm{d}s} - \alpha\left[s\chi(s)^2 + \varphi(s)^2 \right]^{1/3} \varphi(s) = 0 \tag{9a}$$

$$2\frac{\mathrm{d}\varphi(s)}{\mathrm{d}s} + \alpha\left[s\chi(s)^2 + \varphi(s)^2 \right]^{1/3} \chi(s) = 0, \tag{9b}$$

 ISBN: 978-84-608-6082-2

where $\alpha = g\left(\bar{C}C\right)^{1/3}$ for short. Substituting $\chi = As^{-\sigma}F(u)$ and $\varphi = Bs^{-\tau}G(u)$, with $\tau = \frac{3}{4}$, $\sigma = \tau + \frac{1}{2}$, $u = \ln s$ and $A^2 = B^2$ [10],

$$4A|s|^{-\sigma}F(u) - 2s\sigma A|s|^{-\sigma-1}F(u) + 2sA|s|^{-\sigma}F(u)\frac{1}{s} - gB|s|^{-\tau}G(u)\left(sA^2|s|^{-2\sigma}F(u)^2 + B^2|s|^{-2\tau}G(u)^2\right)^{\frac{1}{3}}\left(\bar{C}C\right)^{\frac{1}{3}} = 0 \tag{10a}$$

$$-2\tau B|s|^{-\tau-1}G(u) + 2B|s|^{-\tau}G(u)\frac{1}{s} + gA|s|^{-\sigma}F(u)\left(sA^2|s|^{-2\sigma}F(u)^2 + B^2|s|^{-2\tau}G(u)^2\right)^{\frac{1}{3}}\left(\bar{C}C\right)^{\frac{1}{3}} = 0, \tag{10b}$$

we achieve the dimensionless form of the non-linear ordinary coupled differential equations system (9) as

$$2\frac{\mathrm{d}F(u)}{\mathrm{du}} + \frac{3}{2}F(u) - \alpha\left(AB\right)^{1/3}\left[F(u)^2 + G(u)^2\right]^{1/3}G(u) = 0 \tag{11a}$$

$$2\frac{\mathrm{d}G(u)}{\mathrm{du}} - \frac{3}{2}G(u) + \alpha\left(AB\right)^{1/3}\left[F(u)^2 + G(u)^2\right]^{1/3}F(u) = 0. \tag{11b}$$

Here F and G are dimensionless functions of u and A, B are constants [10]. We call this equations system the Gursey Nonlinear Differential Equations System and the solution of it for $\alpha\left(AB\right)^{1/3} = 1$ is the Gursey Instantons [11]. Gursey Nonlinear Differential Equations System can be written in the form of a vector field as follows by defining a new constant $\gamma \equiv \alpha\left(AB\right)^{1/3}$

$$\boldsymbol{f}_G = \left(-\frac{3}{4}F + \frac{1}{2}\gamma\left[F^2 + G^2\right]^{\frac{1}{3}}G, \frac{3}{4}G - \frac{1}{2}\gamma\left[F^2 + G^2\right]^{\frac{1}{3}}F\right). \tag{12}$$

We take the divergence of $\boldsymbol{f}_G$

$$\nabla.\boldsymbol{f}_G = \left(-\frac{3}{4} + \frac{FG\gamma}{3(F^2+G^2)^{2/3}} + \frac{3}{4} - \frac{FG\gamma}{3(F^2+G^2)^{2/3}}\right) = 0. \tag{13}$$

 ISBN: 978-84-608-6082-2

Density is conserved since $\nabla.\boldsymbol{f}=0$; so the system is conservative. We find the Jacobian of Gursey Nonlinear Differential Equations System as

$$J_{\mathrm{G}}=\begin{pmatrix} -\frac{3}{4}+\frac{FG\gamma}{3(F^2+G^2)^{2/3}} & \frac{G^2\gamma}{3(F^2+G^2)^{2/3}}+\frac{1}{2}(F^2+G^2)^{1/3}\gamma \\ -\frac{F^2\gamma}{3(F^2+G^2)^{2/3}}-\frac{1}{2}(F^2+G^2)^{1/3}\gamma & \frac{3}{4}-\frac{FG\gamma}{3(F^2+G^2)^{2/3}} \end{pmatrix}, \quad (14)$$

and

$$\det(J_G)=-\frac{9}{16}+\frac{FG\gamma}{2(F^2+G^2)^{2/3}}+ \\ \frac{F^2\gamma^2}{6(F^2+G^2)^{1/3}}+\frac{G^2\gamma^2}{6(F^2+G^2)^{1/3}}+\frac{1}{4}(F^2+G^2)^{2/3}\gamma^2. \quad (15)$$

The characteristic eigenvalues providing an understanding of the topological behavior around the singularity points come from $\left|J_{\mathrm{T}}-\lambda I\right|=0$ as

$$\lambda_{\pm}=\pm\frac{\sqrt{27-\frac{24\gamma FG}{(F^2+G^2)^{2/3}}-\frac{8\gamma^2F^2}{(F^2+G^2)^{1/3}}-\frac{8\gamma^2G^2}{(F^2+G^2)^{1/3}}-12\gamma^2(F^2+G^2)^{2/3}}}{4\sqrt{3}}. \quad (16)$$

For $\gamma=0$, one finds reel eigenvalues. Hence the fix point is hyperbolic. For the other γ values, eigenvalues are purely imaginary. So the fix points are elliptic [12].

We redefine Gursey Nonlinear Differential Equations System with the bichromatic potential to get more information about the quantum dynamics of spinor type instantons as

$$2\frac{dF(u)}{du}+\frac{3}{2}F(u)-\gamma\left[F(u)^2+G(u)^2\right]^{\frac{1}{3}}G(u)=0 \quad (17a)$$

$$2\frac{dG(u)}{du}-\frac{3}{2}G(u)+\gamma\left[F(u)^2+G(u)^2\right]^{\frac{1}{3}}F(u)-A_1\cos^2(\omega_1u)+A_2\cos^2(\omega_2u)=0 \quad (17b)$$

$$\frac{dH(u)}{du} = \Omega, \tag{17c}$$

Ω is a constant by adding an extra dimension for numerical calculations. A_1 and A_2 are the amplitudes of external potential and ω_1 and ω_2 are its frequencies respectively.

3. Numerical Results

Different kinds of soliton solutions not only valuable to the application in many fields but also attractive to study the nonlinear phenomena in physics in particular with the aid of the advanced computer technology. Methods from the viewpoint of nonlinear dynamics and chaos theory are quite useful in solving problems where chaos is present. For the Gursey model with potential, it is difficult to obtain exact solutions directly. So we present some numerical results to get more information about the quantum dynamics of spinor type instantons in vacuum. In Figure 1, the stability characterization of Gursey instantons is displayed for $\gamma = 1$ without potential [13,14].

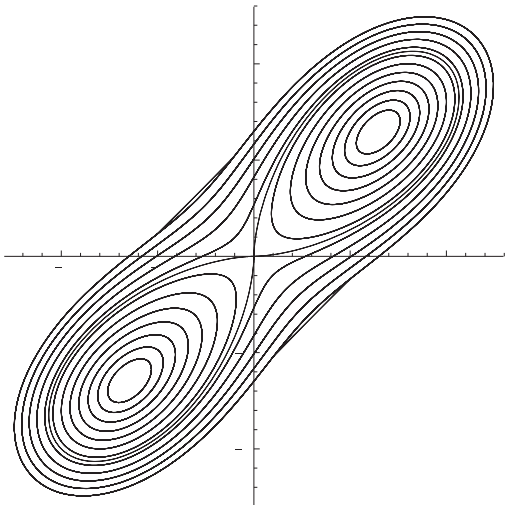
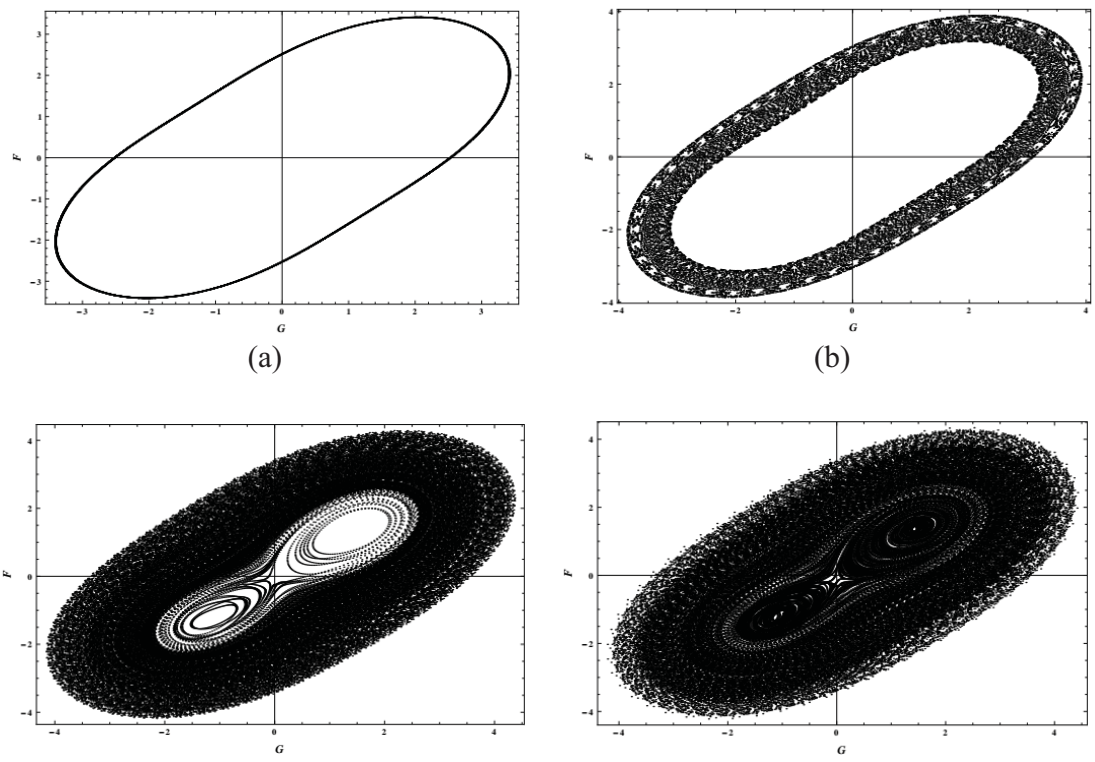


Fig.1: Undamped Duffing type stability characterization of Gursey instantons for $\gamma = 1$; the equilibrium points are $\left(-\frac{3\sqrt{3}}{4}, -\frac{3\sqrt{3}}{4}\right)$ and $\left(\frac{3\sqrt{3}}{4}, \frac{3\sqrt{3}}{4}\right)$.

 ISBN: 978-84-608-6082-2

In Figure 2 the phase space displays for different amplitude values with the initial conditions $F(0) = -\frac{3\sqrt{3}}{4}$, $G(0) = \frac{3\sqrt{3}}{4}$ and $\beta = 1$ are seen.. For the weak potential in Fig. 1(a) the system shows regular behaviour. As we reinforce the amplitude, Fig. 1(b) shows that the chaotic orbits appear in the region near the centre of phase space. Due to more amplified potential, Figs. 3(c) and (d) exhibit more chaotic regions. Hence, one can conclude that external potential having certain frequencies may change the stability characteristics of spinor-type Gursey instantons in phase space for the same initial conditions. When the amplitude of external potential increase enough the Gursey instanton can not pursue its stability for the above initial conditions.



(a) (b)

(a) (b)

Fig.2: Transition to chaos under the bichromatic potential for $\omega_1 = 0.448$ and $\omega_2 = 0.346$ $A_1 = 0.001,\ A_2 = 0.001$ (b) $A_1 = 0.1,\ A_2 = 0.01$ (c) $A_1 = 0.191,\ A_2 = 0.13$ (d) $A_1 = 0.3,\ A_2 = 0.1$

 ISBN: 978-84-608-6082-2

4. Conclusion

In summary, we present the Gursey Nonlinear Differential Equations System formed by the use of Heisenberg ansatz and we investigate it under the bicromatical potential to understand how the behaviours of spinor type Gursey instantons could be affected. The obtained results show vanishing of the stability characteristics of spinor-type Gursey instantons in phase space depending on the external potential parameter values. Instantons lies on the separable closed curves which correspond to regular trajectories without potential. However, potential destroys the separatrix and forms a stochastic layer.

5. References

[1] M. Dunajski, "Solitons, Instantons, and Twistors", Oxford University Press, New York (2010).

[2] C. Rebbi and G. Solliani, Solitons and Particles, World Scientific, (1984); T. Dauoxois and M. Peyrard, Physics of Solitons, Cambridge University Press (2006).

[3] C.Houghton, "Instantons", in the Encyclopaedia of Nonlinear Science, Routledge, NewYork (2005).

[4] R. Rajaraman, Solitons and Instantons, North-Holland, Elsevier Science Publisher, (1982).

[5] F. Aydogmus, "Chaos in a 4D Dissipative Nonlinear Fermionic Model", under review (2014).

[6] F. Gursey, Il Nuovo Cimento 3 (1956), 988-1006.

[7] K.G. Akdeniz and A. Smailagic, "Classical Solitons for Fermionic Models", Il NuovoCimento, **51A** (1979), 345-357.

[8] R. Jackew, Rev. Mod. Phys., 49, 681, (1977); A. CHAKRABARTI, "Introduction to Classical Solutions of Yang-Mills Filed Equations" (1968).

[9] W. Heisenberg, Zeits. F. Naturf. **A9** (1954), 292.

[10] F. Kortel, "On some solutions of Gursey's conformal-invariant spinor wave equation", Il Nuovo Cimento, **4** (1956), 210-215.

[11] K.G. Akdeniz, "On Classical Solutions of Gursey's Conformal-Invariant Spinor Model", Il Nuovo Cimento, **33** (1981), 40-44.

[12] F. Aydogmus and E. Tosyali, Common Behaviours of Spinor Type Instantons in 2-D Thirring and 4-D Gursey Fermionic Models, Advances in High Energy Physics, 148375 (2014), 1-11.

[13] F. Aydogmus, "The Behaviours of Spinor Type Instanton Attractors in Phase Space", Istanbul University, Institute of Science, Physics Department Ph.D. Thesis (2012).

[14] F. Aydogmus, B. Canbaz, C. Onem,K.G. Akdeniz, “The Behaviours of Gursey Instantons in Phase Space, Acta Physica Polonica B, vol.44, **9** (2013), 1837-1845.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engeneering CMMSE 2016

4-8 July, 2016

Surface-induced $L1_0$ ordering processes in nanostructured intermetallics with magnetic anisotropy: Monte Carlo simulation

Sylwia Brodacka[1], Miroslaw Kozlowski[1], Rafal Kozubski[1], Christine Goyhenex[2] and Graeme E. Murch[3]

[1] M. Smoluchowski Institute of Physics, Jagiellonian University in Krakow, Lojasiewicza 11, 30-348 Krakow, Poland

[2] Institut de Physique et Chimie des Materiaux de Strasbourg, Universite de Strasbourg, CNRS UMR 7504, 23 rue du Loess, BP 43,F-67034 Strasbourg, France

[3]The University Centre for Mass and Thermal Transport in Engineering Materials,Priority Research Centre for Geotechnical and Materials Modelling, School of Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia

emails: s.brodacka@gmail.com, kozlowski.miroslaw@gmail.com, rafal.kozubski@uj.edu.pl, christine.goyhenex@ipcms.unistra.fr, Graeme.Murch@newcastle.edu.au

Abstract

The (100)-type surface-induced heterogeneous nucleation of $L1_0$-order domains was observed in a nanolayer, a nanowire and a cubic nanoparticle of FePt. It was found that due to the specific competition between the three kinds of (100)-type free surfaces, the initial c-$L1_0$ variant long-range order appeared to be the most stable in the cubic FePt nanoparticle. The analysis of correlation effects revealed that chemical ordering in initially disordered samples was initiated at the free surfaces.

Key words: chemical ordering, nanostructures, Monte Carlo simulations
*MSC2000:*78M31

 ISBN: 978-84-608-6082-2

1. Introduction.

Free-surface-induced $L1_0$ chemical long-range ordering phenomena in a nanolayer, a nanowire and a cubic nanoparticle of FePt were studied by means of Monte Carlo simulations [1].
Nanostructured $L1_0$ ordered intermetallics such as FePt, FePd, CoPt, have attracted interest for several decades due to their magnetic properties (technology of high density storage media) and to their surface activity (catalysis). The structural origin of the attractive properties of the alloys (first investigated in thin (nano) films and later in nanoparticles) is their $L1_0$ superstructure (Figure 1).

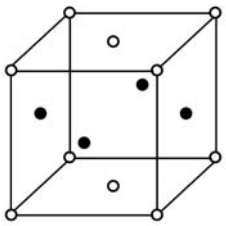


Figure 1 Scheme of the $L1_0$ superstructure of FePt: (●) Fe atoms, (○) Pt atoms. The (100), (010) and (001) orientations of the sequences of alternate Fe and Pt monatomic planes define the a-, b- and c-variants of the $L1_0$ superstructure.

Application of the intermetallics in the technology of high density magnetic storage media depends critically on the preparation of stable nanolayers or nanoparticle matrices showing an *off-plane* direction of the easy magnetization. This means that the material should be $L1_0$ ordered with Fe and Pt monatomic planes (Fig.1) oriented in parallel to the surface.
In our previous works (see [2] and the references therein) we revealed that the (100)-type free surfaces limiting the FePt layers destabilize the $L1_0$ superstructure variants with monatomic crystallographic planes of the same orientation – i.e. parallel to the surface. In view of the technological interest not only in magnetic nanolayers, but also in the matrices of nanoparticles it was important to examine the possible influence of the nanostructure geometry on the stability of the homogeneous $L1_0$ superstructure in FePt.

2. Methodology

The system was modeled with nearest-neighbor and next-nearest-neighbor interatomic pair interactions deduced from ab initio calculations. The generated

 ISBN: 978-84-608-6082-2

samples, the dimensionality of which was determined by appropriate periodic boundary conditions imposed upon the generated supercells, were initially either perfectly ordered in the c-variant $L1_0$ superstructure (Fig.1), or completely disordered in the fcc crystalline structure. Vacancy-mediated creation of equilibrium atomic configurations was modelled by relaxing the systems at temperatures below the 'order-disorder' transition point using the Glauber algorithm implemented with the vacancy mechanism of atomic migration.

3. Results

The (100)-type-surface-induced heterogeneous nucleation of $L1_0$-order domains was observed (Fig.2) and quantified by means of an original parameterization enabling selective determination of volume fractions of particular $L1_0$- variants.
Due to the specific competition between the three kinds of (100)-type free surfaces, the initial long-range order in the c-$L1_0$ variant appeared to be the most stable in the cubic nanoparticle. The free surfaces limiting the nanocube effectively stabilized the initial superstructure. The occurrence of this phenomenon is one of the most important conclusions of the present study.

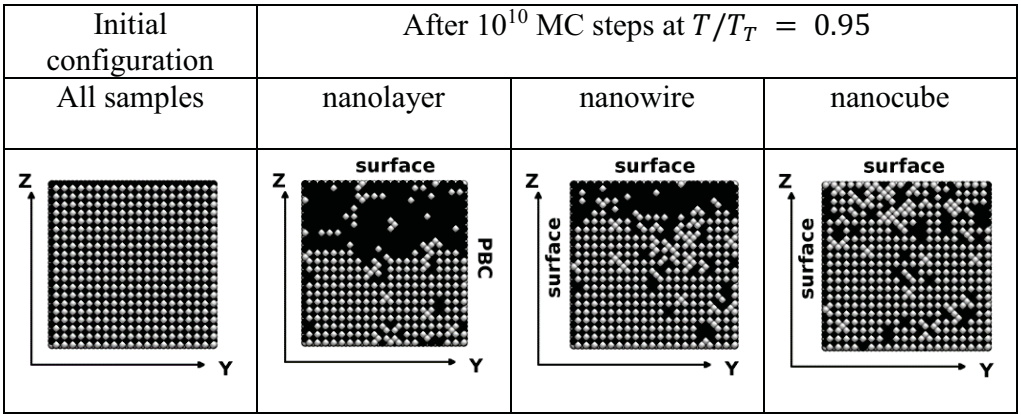


Figure 2 Atomic configurations in the samples initially ordered in the c-$L1_0$ variant and annealed for 10^{10} MC steps at $T/T_T = 0.95$ (T_T is the "order-disorder" transition point). Black and white dots represent Fe and Pt atoms, respectively.

Atomic ordering starting from initially disordered FePt samples was simulated exclusively at $T/T_T = 0.95$. Images of atomic configurations generated in the samples by 10^{10} MC steps (Fig.3) show well-marked $L1_0$-variant domain structures with statistically distributed antisite defects.

 ISBN: 978-84-608-6082-2

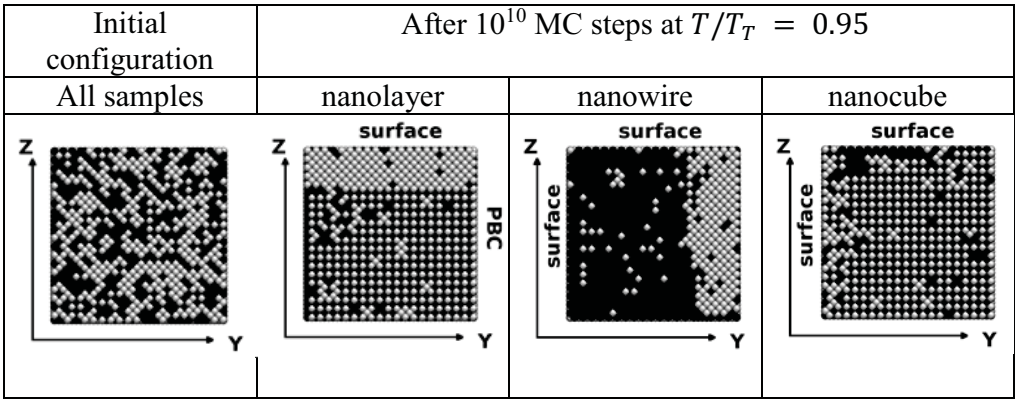


Figure 3 Atomic configurations in the samples initially disordered and annealed for 10^{10} MC steps at $T/T_T = 0.95$ (T_T is the "order-disorder" transition point). Black and white dots represent Fe and Pt atoms, respectively.

While almost purely a-$L1_0$ variant superstructure was observed in the [100]-oriented nanowire, a mosaic of a- and b-$L1_0$ variant domains was generated in the

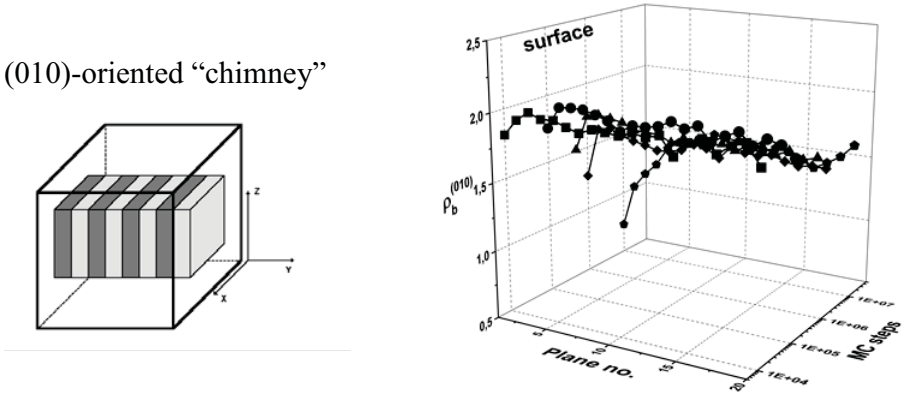


Figure 4 Pair-correlation parameter $\rho_b^{(010)}$ in initially disordered single FePt nanocube against the numbers of consecutive (010)-planes in the [010]-oriented 'chimney'. (■) 10^4 MC steps, (▲) 10^5 MC steps, (●) 5×10^5 MC steps, (▼) 10^6 MC steps, (⬟) 10^7 MC steps.

(001)-oriented nanolayer. No definite selection of the $L1_0$-variants in the nanocube resulted, in turn, in almost monovariant LRO in the inner part and a quite high contribution of all other $L1_0$ variants close to the surface. The latter was inevitable due to any variant winning inside the cube being always unfavourable for one pair of surfaces. The analysis of correlation effects performed by means of a parameter

very sensitive to any deviation from the statistical distribution of atoms revealed that chemical ordering was initiated at the free surfaces (Fig.4).

Acknowledgments

Support is acknowledged from the COST Action MP0903 NANOALLOY. Two of the authors (R.K. and G.E.M.) acknowledge support from the European Community's Seventh Framework Programme (FP7-PEOPLE-2013-IRSES) under *EC-GA no.* 612552 and from the funds of the Polish Ministry of Science and High Education (Grant no. 3135/7. PR/2014/2). Access to the computation facilities at the ACK Cyfronet AGH, Krakow (supercomputer ‘‘Zeus’’) is greatly appreciated. The authors are also grateful to Dr. Lukasz Zosiak for the valuable assistance at the numerical calculations.
The figures are reproduced from Ref.[1] with permission from the PCCP Owner Societies.

4. References

[1] S. BRODACKA, M. KOZLOWSKI, R. KOZUBSKI, CH. GOYHENEX, G.E. MURCH, *Chemical ordering phenomena in nanostructured FePt: Monte Carlo Simulations,* Phys. Chem. Chem. Phys. **17** (2015) 28394 – 28406.

[2] M. KOZLOWSKI, R. KOZUBSKI, C. GOYHENEX, *Surface induced superstructure transformation in* $L1_0$ FePt *by Monte Carlo simulations implemented with Analytic Bond-Order Potentials*, Materials Letters **106** (2013) 273–276

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Advancing Algorithms to Increase Performance of Correlated and Dynamical Electronic Structure Simulations

Wibe A. de Jong[1], Mathias Jacquelin[1], Eric J. Bylaska[2], Konstantinos Vogiatzis[3], Laura Gagliardi[3]

[1] *Computational Research Division, Lawrence Berkeley National Laboratory*
[2] *Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory*
[3] *Department of Chemistry, University of Minnesota*

emails: wadejong@lbl.gov, mjacquelin@lbl.gov, eric.bylaska@pnnl.gov, kvogiatz@umn.edu, gagliard@umn.edu

Abstract

Modelling strongly correlated systems and systems in dynamical environments requires fast, computationally scalable and efficient computational tools utilizing novel mathematical approaches and the latest hardware technologies. Parallel scaling of a newly developed parallel MCSCF code with NWChem to tens of thousands of processors and active spaces beyond CAS(20,20) is demonstrated. The Intel Phi multicore architecture combined with algorithmic advances were used achieve 5.5x improved performance for key kernels in plane wave *ab initio* molecular dynamics.

Key words: MCSCF, Plane Wave Dynamics, Parallel, Intel Phi

1. Introduction

The US Department of Energy is focusing a significant fraction of its resources to research on controlling chemical and physical processes in dynamical environments at the molecular and nanoscale that can lead to new or more efficient renewable energy resources (e.g. energy storage materials such as batteries, light harvesting systems, catalysts) and approaches to reduce the carbon footprint. Many relevant processes are driven by complex electronic structure

 ISBN: 978-84-608-6082-2

transitions and strongly correlated electronic systems that can be described by multi-configurational self-consistent field (MCSCF) methods [1]. To date, practical applications of this method has been restricted to studies of 20 electrons in 20 orbitals and takes huge amounts of wall-clock time on a single node. To achieve the fastest time-to-solution for the MCSCF, and to enable access to larger orbitals and electron spaces, computationally parallel, scalable and efficient computational algorithms need to be developed. In addition, most chemical reactions happen in complex and dynamical environments, often away from equilibrium. One methodology to describe these processes is a*b initio* molecular dynamics, for example plane wave density functional theory based *ab initio* molecular dynamics. These methods are computationally expensive, and often limited to tens of picoseconds and hundreds of atoms. To enable simulations of longer time and length scales, algorithms need to be modified and optimized to take fully take advantage of the latest hardware technologies available to the computational chemistry community.

In this paper we will discuss progress made in the development of a parallel MCSCF code. In the second part we will discuss our efforts to accelerate the solution of plane wave *ab initio* molecular dynamics algorithms through algorithmic developments utilizing the Intel Phi multicore architectures.

2. Parallel MCSCF enables record size simulations

A new parallel MCSCF code has been developed within the open-source NWChem software suite [2]. In an MCSCF, the most time consuming step the the configuration interaction (CI) algorithm, and specifically the calculation of the so-called sigma vector. An algorithm was devised that distributes the large CI vector

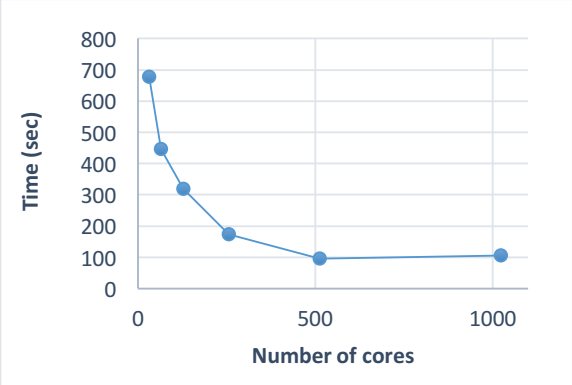


Figure 1: Scalability results for the Cr_3 CAS(20,20) molecular system.

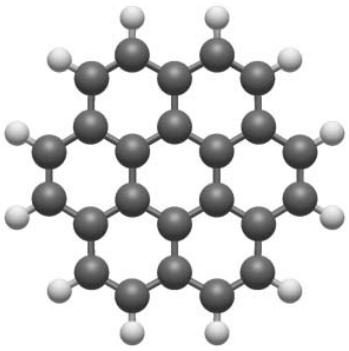


Figure 2: Corenene molecular system.

of Slater Determinants over the available processors, while two-electron integrals are kept local to minimize bottlenecks related to data communication. A novel load-balancing scheme was developed, utilizing the general active space (GAS)

 ISBN: 978-84-608-6082-2

approach to spit the CI vector in smaller blocks, and distribute the amount of work per processor based on the length of the CI vector. Additional optimizations are ongoing where in addition to CI vectors length distribution the scheme will also approximate the level of work each section of the CI vector will be involved in. The parallel MCSCF algorithm was tested on the NERSC Cori machine and the Cr_3 CASCI with 20 electrons in 20 orbitals leading to 2 billion Slater Determinants (SD), a calculation that is considered the largest possible and is extremely difficult to achieve with currently available codes, can be done in a matter of minutes on 500-1000 processors (see figure 1).

To explore the limits of the new MCSCF code in NWChem, initial benchmark calculations were performed on a molecular system with 24 electrons in 24 orbitals, a so-called CAS(24,24) with a half a trillion Slater Determinants (3.6 trillion SDs without symmetry), a record calculation that far exceeds what has been currently feasible for the scientific community. NWChem's new highly scalable MCSCF algorithm can run a single CI iteration corresponding on 7680 processors in 15 minutes. Scalability testing, and additional load-balancing improvements to increase the number of processors to get an even faster time-to-solution are ongoing. With the ability to tackle problem sizes of CAS(24,24) and beyond, NWChem provides access to scientific problem sizes inaccessible until now. A good example are extended π-conjugated systems, such as coronene (see figure 2), relevant to photochemistry and building functional nanodevices.

3. Improving performance of plane wave codes on Intel Phi

Many-core CPU architectures, such as Intel's Phi, will become the main stream resources for high-performance scientific computing. For example, LBNL's NERSC Supercomputing Center will be taking delivery of their next petaflop system consisting of Intel Knight's Landing Processors in 2016. Algorithms need to be modified and optimized to fully take advantage of a large number of hardware cores and threads per compute node. Here the focus is to optimize the performance of the Lagrange multiplier kernel in the plane wave ab initio molecular dynamics code in NWChem. This particular kernel is used when orthogonalizing the wave function, and consists of a sequence of dense matrix-matrix products. The wave function is represented by a n_{pack}-*by*-n_e matrix, where n_e represents the number of electrons and n_{pack} is the number of plane wave basis functions. The particularity of the Lagrange multiplier kernel is that n_e is significantly smaller than n_{pack}. Three types of matrix products need to be performed for each wave function during the computation:

1. A matrix with *ne* rows and n_{pack} columns is first multiplied by a matrix with n_{pack} rows and *ne* columns, therefore resembling an *inner product*.
2. Two n_e-by-n_e matrices are multiplied together.
3. A matrix with n_{pack} rows and *ne* columns is multiplied by a n_e-by-n_e matrix.

 ISBN: 978-84-608-6082-2

Our experiments show that current BLAS implementations do not efficiently exploit the available parallelism in the first type of matrix product (see figure 3). Work was performed to develop an algorithm that better leverages the processing power of many-core processors on this specific *inner matrix product*, which is frequently encountered in computational chemistry codes.

By replicating the n_e-by-n_e resulting matrix onto each thread, then distributing the work over the larger dimension (n_{pack}), and finally reducing each thread's contribution into the final output matrix, our implementation achieves a significantly higher parallel efficiency on Intel Knight's Corner processors. Moreover, as the Lagrange multiplier kernel handles three sets of wave functions, it is important to limit the number of synchronization points so that these three independent operations can be processed seamlessly. Our experiments show that best performance is achieved when using a combination of 4 MPI processes and 40 threads per process, a full Lagrange multiplier kernel execution can be performed in less than 0.17s when n_{pack} = 18008 and a number of electrons n_e = 256.

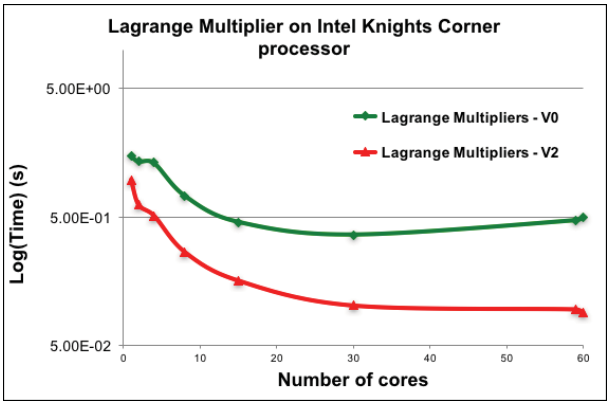


Figure 3: Comparison of the standard BLAS Intel MKL library (V0) and new reduction based (V2) algorithm performance of the Lagrange multiplier for an n_{pack}-by-n_e of 5000 by 200. A 5.5x performance was achieved.

Overall, the results show the importance of preparing computational chemistry codes to the upcoming generation of processors that expose an unprecedented level of parallelism within each socket. A traditional approach solely relying on multithreaded BLAS operations is very unlikely to successfully exploit the available parallelism. It is therefore critical to express as much parallelism at the algorithm level and to remove synchronization points.

4. Acknowledgements

Authors of this work were supported by the DOE Office of Advanced Scientific Computing Research (ASCR) under contract number DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC). This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

5. References

[1] B.O. ROOS, P.R. TAYLOR AND P.E.M. SIEGBAHN, *A Complete Active Space SCF Method (CASSCF)*, Chem. Phys. **48** (1980) 157-173.

[2] M. VALIEV, E.J. BYLASKA, N. GOVIND, K. KOWALSKI, T.P. STRAATSMA, H.J.J. VAN DAM, D. WANG, J. NIEPLOCHA, E. APRA, T.L. WINDUS AND W.A. DE JONG, *NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations*, Comput. Phys. Commun. **181** (2010), 1477.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Piecewise Modelling and Simulation of a Rotating Extensible Manipulator Link for Base Placement and Path Smoothness

Mihai Dupac[1]

[1] *Department of Design and Engineering, Bournemouth University*

email: mdupac@bournemouth.ac.uk

Abstract

Optimal trajectory planning plays an important role in many industrial applications where robotic manipulators do not provide the required productivity. Since productivity is affected by the manipulator speed, its end effector trajectory should take into consideration the manipulator dynamic characteristics such as the minimization of mechanical energy, minimization of link deflections and actuator constraints. To achieve these requirements and complete manipulator design, new motion planning algorithms - based on accurate mathematical modelling and simulation of the manipulators trajectories - should be developed.

In this paper the analysis and conversion of the polygonal trajectory of the end-effector of a rotating extensible link manipulator to a curved trajectory (passing through each via point) using piecewise polynomial interpolation is considered. The polygonal trajectory is modelled using cubic curves in order to generate a smooth path and minimize link deflection - caused by link flexibility, clearance and deformation. In this context, the possible placement of the robotic manipulator base was examined using computational geometry techniques and related to its geometric path.

Key words: template, instructions
MSC2000: AMS Codes (optional)

1. Introduction

Optimal trajectory planning plays an important role in a wide range of industrial applications where robotic manipulators are used for improving productivity and

 ISBN: 978-84-608-6082-2

reducing production costs. To obtain the desired requirements, some of the existing algorithms [20, 22, 24] should be revised and new planning algorithms should be developed while considering the complexity of the new robotic arm design [13,16,17] and new requirements such as collision avoidance and joint characteristics [6,10,24].

The increase complexity of the new manipulators like the newly developed medical robots, require rigorous trajectory planning studies [2,3]. Some manipulator requirements involving optimal trajectory planning while performing a specific task, reside in finding the "most suitable" location/orientation of the manipulator base [1,10,25]. Due to the manipulator dimensions and complexity, a proper method for finding an optimal location of the manipulator base is still an open problem (and will be part of this study).

Generating trajectories – while considering some desired features and objectives - represents key issues in robotic applications. Path planning with automatic obstacle avoidance including kinematics and dynamics constraints [15], execution time [7, 8] and jerk [11] are just some of the important features in trajectory planning. Continuous trajectories usually represented by piecewise interpolating curves with slope continuity, geometrically continuous Catmull-Rom splines [9], parametric and/or geometric continuous splines [21, 31] or uniform Cubic B-Spline with parametric and geometric continuity, are adequate tools in generating a smooth motion of the manipulator especially when manipulator dynamics is considered. In this context, a geometric approach was considered in [10] to determine the base location for simple manipulators configurations, while an optimal location to address manipulator base location considering velocity performance was discussed in [26].

In this paper the trajectory planning of the end-effector (smooth path which minimize link deflection caused by clearance and deformation of flexible components) of a rotating extensible robotic arm/link manipulator is considered. The desired geometric path of the end-effector is generated by converting of the polygonal trajectory of the end-effector to a curved trajectory using piecewise polynomial interpolation. A 'most appropriate' placement of the base of the robotic manipulator is examined and related to its geometric path.

2. Mathematical Modelling

Manipulator Model

For a trajectory planning study of the rotating extensible manipulator arm, the mechanical model composed of a rigid link and a sliding (Fig. 1) is considered. The end effector of the manipulator follow the points $\mathbf{p}_0\mathbf{p}_1...\mathbf{p}_i...\mathbf{p}_n$ in Fig. 1.

 ISBN: 978-84-608-6082-2

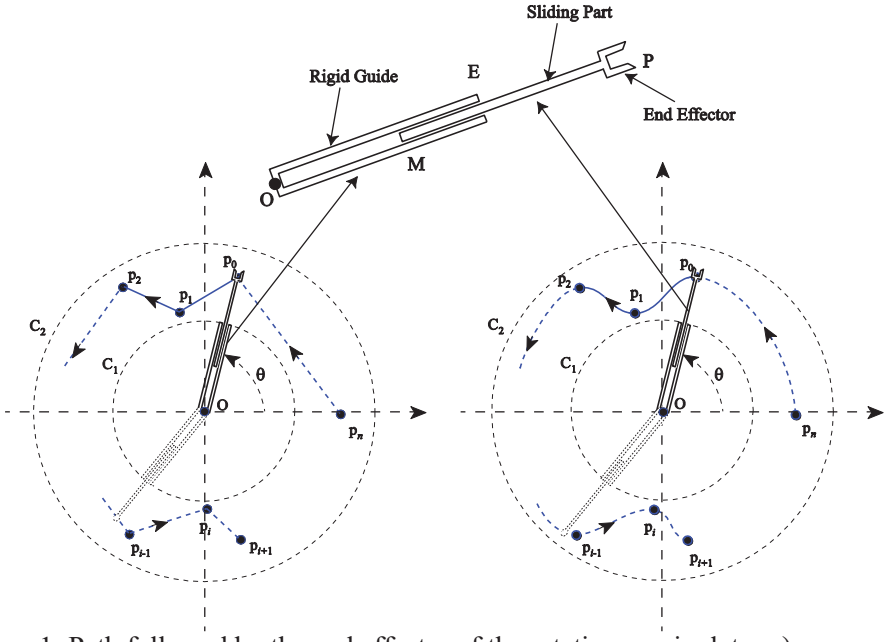


Figure 1: Path followed by the end effector of the rotating manipulator a) non convex trajectory, b) non-convex C^2 continuous parametric trajectory

The rigid guide of the manipulator denoted by OE has length d_{OE}, the sliding part which is denoted by SP - with its non-active end denoted by S and its active end-effector denoted by P- has the length d_{SP}. The distance d_{OP} between the manipulator base location O and its end-effector P varies due to the motion (rotation) of the rigid guide of the extensible manipulator. The manipulator can perform quick stops after a θ degree rotation, where θ represents the angle between the guide and the horizontal direction.

Piecewise Trajectory Generation

In order to achieve some desired properties of the manipulator trajectory piecewise polynomial interpolating curves may be considered. Such interpolating curves [7,8,15,19,21,26] guarantee slope continuity, and/or minimal data storage, and/or local control and smoothness (no abrupt changes in displacement and velocity). Good examples are the cubic splines (Hermite, Cardinal, Catmul-Rom) that can be represented by $\mathbf{p}(u)=\mathbf{a}_0+u\mathbf{a}_1+u^2\mathbf{a}_2+u^3\mathbf{a}_3$ where $\mathbf{a}_0=(a_{0x},a_{0y},a_{0z})$, $\mathbf{a}_1=(a_{1x},a_{1y},a_{1z}),\mathbf{a}_2=(a_{2x},a_{2y},a_{2z}),\ \mathbf{a}_3=(a_{3x},a_{3y},a_{3z})$.

The tangent and the normal to such a curve is represented using the unit tangent vector and unit normal vector $\mathbf{n}(u)$ defined by $\mathbf{t}(u)=\dfrac{\dot{\mathbf{p}}(u)}{\|\mathbf{p}(u)\|}$, $\mathbf{n}(u)=\dfrac{\dot{\mathbf{t}}(u)}{\|\mathbf{t}(u)\|}$.

A Hermite cubic spline is defined by two successive control points (position) along with the slope continuity at each control point (1st derivative of endpoints).

 ISBN: 978-84-608-6082-2

A Hermite cubic spline is represented by a cubic polynomial, defined as

$$\begin{aligned}\mathbf{p}(u) &= \mathbf{a}_0 + u\mathbf{a}_1 + u^2\mathbf{a}_2 + u^3\mathbf{a}_3 \\ &= \mathbf{p}_0 + u\mathbf{p}_1 + u^2\left(-3\mathbf{p}_0 - 2\mathbf{p}_1 + 3\mathbf{p}_2 - \mathbf{p}_3\right) + u^3\left(2\mathbf{p}_0 + \mathbf{p}_1 - 2\mathbf{p}_2 + \mathbf{p}_3\right) \\ &= \left(2u^3 - 3u^2 + 1\right)\mathbf{p}_0 + \left(u^3 - 2u^2 + u\right)\mathbf{p}_1 + \left(-2u^3 + 3u^2\right)\mathbf{p}_2 + \left(u^3 - u^2\right)\mathbf{p}_3\end{aligned} \quad (1)$$

where the control points $\mathbf{p}_0$ and $\mathbf{p}_2$, as well as the control slopes and are defined by $\mathbf{p}_0 = \mathbf{p}(0) = \mathbf{a}_0$, $\mathbf{p}_1 = \dot{\mathbf{p}}(0) = \mathbf{a}_1$, $\mathbf{p}_2 = \mathbf{p}(1) = \mathbf{a}_0 + \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3$, $\mathbf{p}_3 = \dot{\mathbf{p}}(1) = \mathbf{a}_1 + 2\mathbf{a}_2 + 3\mathbf{a}_3$.

A Cardinal cubic spline is defined by two successive control points (position) $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, along with the slopes at the control points which are determined from previous and subsequent control points $\mathbf{p}_{i-1}$ and $\mathbf{p}_{i+2}$ in the sequence (the derivatives at $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ are determined by the vectors $\mathbf{p}_{i+1} - \mathbf{p}_{i-1}$ and $\mathbf{p}_{i+2} - \mathbf{p}_i$ respectively). A Cardinal cubic spline with tension is represented in its canonical form by:

$$\begin{aligned}\mathbf{p}(u) &= \mathbf{a}_0 + u\mathbf{a}_1 + u^2\mathbf{a}_2 + u^3\mathbf{a}_3 \\ &= \mathbf{p}_1 + us(\mathbf{p}_2 - \mathbf{p}_0) + u^2\left[2s\mathbf{p}_0 + \mathbf{p}_1(s-3) + \mathbf{p}_2(3-2s) - s\mathbf{p}_3\right] \\ &\quad + u^3\left[-s\mathbf{p}_0 + \mathbf{p}_1(2-s) + \mathbf{p}_2(s-2) - s\mathbf{p}_3\right]\end{aligned} \quad (2)$$

where the control points are defined by $\mathbf{p}_1 = \mathbf{p}(0) = \mathbf{a}_0$, $\mathbf{p}_2 = \mathbf{p}(1) = \mathbf{a}_0 + \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3$, $\dot{\mathbf{p}}(0) = \mathbf{a}_1$, $\dot{\mathbf{p}}(1) = \mathbf{a}_1 + 2\mathbf{a}_2 + 3\mathbf{a}_3$ and where the tension $t = 1 - 2s$ adjusts how much the derivatives are scaled, that is, $\dot{\mathbf{p}}(0) = s(\mathbf{p}_2 - \mathbf{p}_0)$, $\dot{\mathbf{p}}(1) = s(\mathbf{p}_3 - \mathbf{p}_1)$. Positive tension values tighten the trajectory while negatives tension values loosen the trajectory. A particular case of cardinal splines are the Catmul-Rom splines having the tension value t = 0.

Since the cubic splines mentioned above lose their second-order continuity, uniform Cubic B-Spline (having C^2 parametric continuity – motion related, and G^2 geometric continuity – design related) should be used instead. The use of Catmull-Rom Splines [9] or uniform Cubic B-Spline having C^2 and G^2 continuity [4] are necessary when the uniqueness of the normal (vector normal to the curve) is needed, that is, the curve has a constant curvature. A well approached example of geometric continuous cubic B-Spline have been considered in [4] by

$$\mathbf{p_i}(u) = \mathbf{a}_0(\beta_1, \beta_2) + u\mathbf{a}_1(\beta_1, \beta_2) + u^2\mathbf{a}_2(\beta_1, \beta_2) + u^3\mathbf{a}_3(\beta_1, \beta_2) \quad (3)$$

where $\mathbf{a}_0(\beta_1, \beta_2), \mathbf{a}_1(\beta_1, \beta_2), \mathbf{a}_2(\beta_1, \beta_2)$ and $\mathbf{a}_3(\beta_1, \beta_2)$ are cubic polynomial functions constructed so that $\mathbf{p_{i+1}}(0) = \mathbf{p_i}(1)$, $\dot{\mathbf{p}}_{\mathbf{i+1}}(0) = \beta_{1i}\dot{\mathbf{p}}_{\mathbf{i}}(1)$, $\ddot{\mathbf{p}}_{\mathbf{i+1}}(0) = \beta_{1i}^2\ddot{\mathbf{p}}_{\mathbf{i}}(1) + \beta_{2i}\dot{\mathbf{p}}_{\mathbf{i}}(1)$. The cubic B-Spline described above has the desired C^2 continuity since "two curves meet with G^2 continuity if and only if their arc length parameterization meet with C^2 continuity"[9].

 ISBN: 978-84-608-6082-2

3. Computation of Base Location and Trajectory Generation

The design of a manipulator trajectory in which the end effector can move smoothly along a desired path can be obtained using piecewise continuous parametric curves which can accurately describe and control the shape. Since for most of manipulator tasks the end-effector should actually pass through the points while providing continuity of the manipulator trajectory positions, velocities and accelerations, the use of such interpolating curves is required.

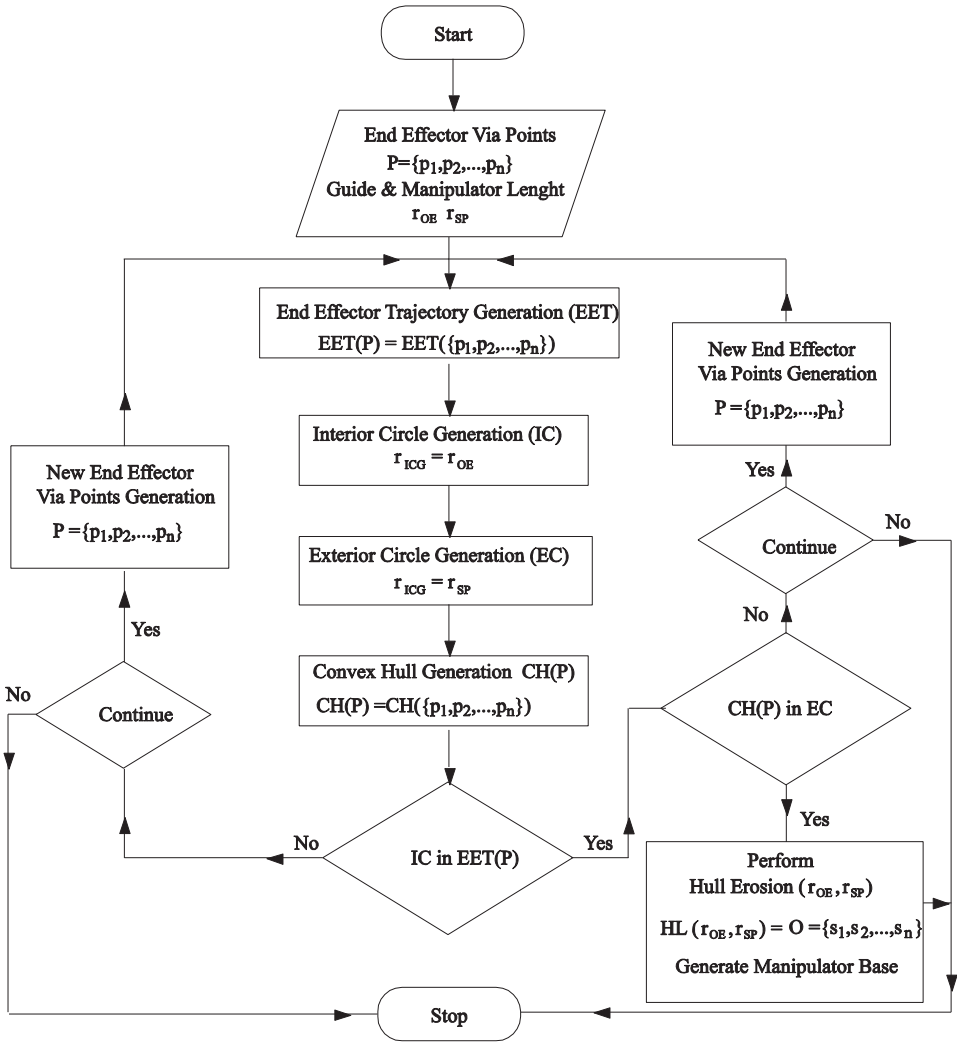


Figure 2. Base Location algorithm

 ISBN: 978-84-608-6082-2

Due to the (relative) easiness in finding a possible placement of the robotic manipulator base when its end-effector follows a convex trajectory, the design of a convex and a non-convex trajectory and associated base location is treated separately using computational geometry techniques [5,12].

Convex Case

The first step in determining the manipulator base location is to find if such a location exists. If polygonal trajectory of the end-effector of the manipulator is convex, a location of the manipulator base exists if and only if the system

$$\begin{cases} d_{OP_i} \le d_{OE} + d_{SP} \\ \max(d_{OE}, d_{SP}) \le d_{OP_i} \end{cases} \tag{4}$$

has a solution. To determine a possible location of the manipulator base - for either a convex polygonal trajectory or convex piecewise trajectory - the Base Location algorithm described in Fig. 2 should be applied.

Non-Convex Case

The first step in determining the manipulator base location - when the trajectory of the end effector is non-convex - is to find if such a location exists. In this case, if the number of guards obtained using the Art Gallery Problem (AGP) [12,14,18] algorithm is greater than one, the manipulator base problem has no solution, otherwise (number of guards equal to one) a location of the manipulator base exist if and only if Eq. (4) has a solution. Since Eq. (4) fails (circle C_1 cannot be inscribed inside the non-convex polygonal curve or non-convex piecewise trajectory shown in Fig. 3a and Fig. 3b respectively) the trajectory shown in Fig. 3 does not have a solution (although AGP is verified), and therefore a location of the manipulator base cannot be determined.

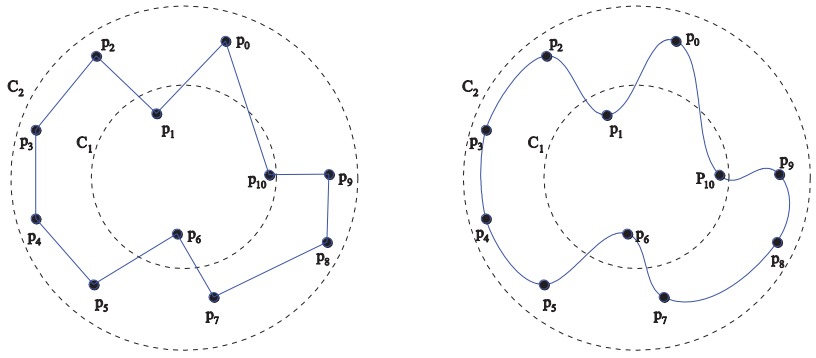


Figure 3: Manipulator (a) non-convex polygonal trajectory, and (b) non-convex C^2 continuous parametric trajectory

 ISBN: 978-84-608-6082-2

4. Results

Two examples are presented to illustrate the placement of the base and trajectory planning of the end-effector of the rotating extensible link for both convex and non-convex trajectories. Simulations have been performed [23] for a manipulator composed of a rigid link of length d_{OE} = 0.35 m and a sliding link of length d_{SP} = 30 cm.

Convex Case

Using the via points $\mathbf{p}_0$, $\mathbf{p}_1$, …., $\mathbf{p}_{10}$ of the convex polygonal trajectory in Fig. 4a a new manipulator trajectory [4,9] in which the end effector can move smoothly along the path have been obtained (Fig. 4b). The possible locations of the manipulator base for the polygonal trajectory (Fig. 4a) and pricewise trajectory (Fig. 4b) have been determined using the Base Location algorithm. The location of the manipulator base of the polygonal and pricewise trajectory is the interior of the green path shown in Fig. 4a and Fig. 4b respectively.

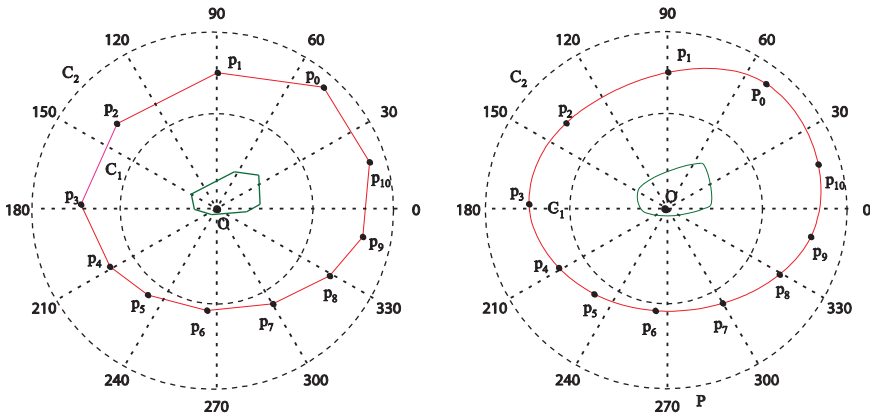


Figure 4: (a) convex polygonal trajectory (red) and base location (green), (b) convex piecewise trajectory (red) end and base location (green)

Non-Convex Case

The AGP algorithm was applied for the particular trajectory described by the points $\mathbf{p}_0$, $\mathbf{p}_1$, …., $\mathbf{p}_8$ in Fig. 5a and Fig. 5b.

 ISBN: 978-84-608-6082-2

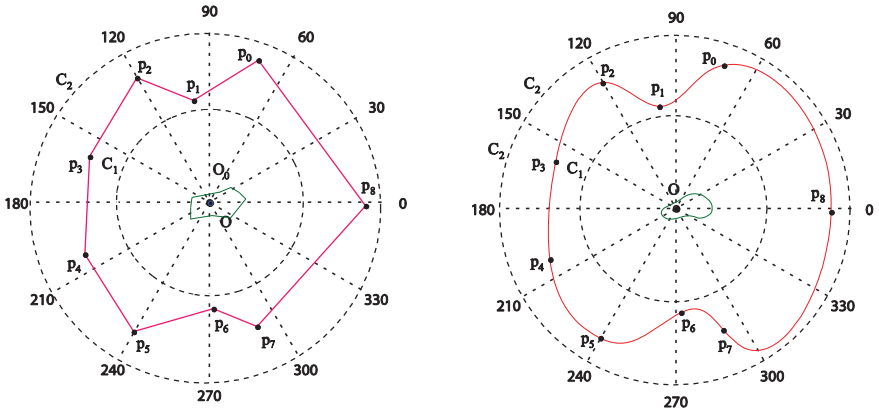


Figure 5: (a) non-convex polygonal trajectory (red) and base location (gren), (b) non-convex piecewise trajectory (red) and base location (green)

In this case one guard solve the visibility problem for both polygonal and its piecewise interpolating curve (Fig. 5a and Fig. 5b) which indicate a possible solution for the manipulator base location in both cases. A final validation and location of the manipulator base obtained using Eq. (4) is represented by the interior of the green path shown in Fig. 5a.

For the modified curved trajectory of the end-effector of the manipulator (Fig. 5b), the possible base location determined using the Base Location algorithm has as graphical representation the interior the green piecewisee parametric curve shown in Fig. 5b.

5. Conclusions

In this paper the conversion of a polygonal trajectory of the end-effector of a rotating extensible robotic arm manipulator to a smooth trajectory - which minimize link deflection caused by clearance and deformation of flexible components - using piecewise parametric curves is presented. The placement of the robotic manipulator base is examined and related to its continuous geometric path using cubic spline interpolating curves and computational geometry techniques such as convex hull, largest empty circle problem, smallest enclosing circle, triangulation and vertices 3-coluring. Simulations have been performed for both the convex and non-convex case to show the newly generated trajectories

6. References

 ISBN: 978-84-608-6082-2

[1] K. ABDEL-MALEK, W. YU, J. YANG, K. NEBEL, *A mathematical method for ergonomic-based design: placement*, International Journal of Industrial Ergonomics, **34(5)**, 2004, 375-394

[2] L. ADHAMI, E. COSTE-, *Optimal planning for minimally invasive surgical robots*, IEEE Transactions on Robotics and Automation, **19(5)**, 854–863, 2003

[3] MANIERE L. ZHAO, A. JOUBAIR, P. BIGRAS, I.A. BONEV, *Metrological evaluation of a novel medical robot and its kinematic claibration*, International Journal of Advanced Robotic Systems, **12:126**, 2015

[4] B.B. BARSKY, T.D. DEROSE, *Three Characterizations of Geometric Continuity for Parametric Curve*, University of California at Berkeley Technical Report No. UCB/CSD 88/417, 1988

[5] M. DE BERG, M. VAN KREVELD, M. OVERMARS, O. SCHWARZKOPT, *Computational Geometry, Algorithms and Applications*, Springer, 2000

[6] H. DONG, Z. DU, *Obstacle avoidance path planning of planar redundant manipulators using workspace density*, International Journal of Advanced Robotic Systems, **12(9)**, 2015

[7] J.E. BOBROW, S. DUBOWSKY, J.S. GIBSON, *Time-optimal control of robotic manipulators along specified paths*, Int. J. Robot. Res., **4(3)**, 554–561, 1985

[8] D. CONSTANTINESCU, E.A. CROFT, *Smooth and time-optimal trajectory planning for industrial manipulators along specified paths*, J. Robot. Syst, **17(5)**, 233–249, 2000

[9] T.D. DEROSE, B. A. BARSKY, *Geometric Continuity, Shape Parameters, and Geometric Constructions for Catmull-Rom Splines*, ACM Transactions on Graphics, 7(1), 1-41, 1988.

[10] J.J. YANG, W. YU, J. KIM, K. ABDEL-MALEK, *On the placement of open-loop robotic manipulators for reachability*, Mechanism and Machine Theory, **44**, 671–684, 2009

[11] J. DONG, P.M. FERREIRA, J.A. STORI, *Feed-rate optimization with jerk constraints for generating minimum-time trajectories*, Int. J. Mach. Tools Manuf., **47(12)**, 1941–1955, 2007

[12] M. DUPAC, *Trajectory Planning for a Rotating Extensible Robotic Manipulator: Design for Collision Avoidance and Accurate Position Placement*, Proc. of the 9th International Conference on Engineering Computational Technology", Paper 56, 2014.

[13] S.K. DWIVEDY, P. EBERHARD, *Dynamic analysis of flexible manipulators, a literature review*, Mechanism and Machine Theory **41**, 749–777, 2006

[14] S.K. GHOSH, *Visibility algorithms in the plane*, Cambridge Univ. Press, 2007

[15] D. HSU, R. KINDEL, J.C. LATOMBE, S. ROCK, *Randomized kinodynamic motion planning with moving obstacles*, Int. J. Robot. Res., **21(3)**, 233–255, 2002

[16] M. DUPAC, *Dynamical analysis of a constrained flexible extensible link*

with rigid support and clearance, Journal of Theoretical and Applied Mechanics **52(3)**, 665-676, 2014.

[17]M. KALYONCU, *Mathematical modelling and dynamic response of a multistraight-line path tracing flexible robot manipulator with rotating-prismatic joint*, Applied Mathematical Modelling 32, 1087–1098, 2008

[18]J. O'ROURKE, *Art Gallery Theorems and Algorithms*, Oxford University Press, 1987

[19]S.-R. CHANG, U.-Y. HUH, *A collision-free G^2 continuous path-smoothing algorithm using quadratic polynomial interpolation*, International Journal of Advanced Robotic Systems, **11(194)**, 2014

[20]Z. KONJOVIC, M. VUKOBRATOVIC, D. SURLA, *Synthesis of the optimal trajectories for robotic manipulators based on their complete dynamic models*, Int. J. Robotics Autom., **9(1)**, 36–47, 1994.

[21]M. KRAJNC, *Interpolation Scheme for Planar Cubic G2 Spline Curves*, Acta Applicandae Mathematicae, **113(2),** 129-143, 2011

[22]M.C. LEU, S.K. SINGH, *Optimal planning of trajectories robots", in "CAD Based Programming for Sensory Robots*, B. Ravani (Editor), Springer, 1988.

[23]D.B. MARGHITU, M. DUPAC, *Advanced Dynamics: Analytical and Numerical Calculations with MATLAB*, Springer, 2012

[24]S. MITSI, K.-D. BOUZAKIS, G. MANSOUR, *Optimization of robot links motion in inverse kinematics solution considering collision avoidance and joint limits*, Mech. Mach. Theory, **30(5)**, 653–663, 1995

[25]S. MITSI, K.-D. BOUZAKIS, D. SAGRIS, G. MANSOUR, *Determination of optimum robot base location considering discrete end-effector positions by means of hybrid genetic algorithm*, Robotics and Computer-Integrated Manufacturing, **24(1)**, 50-59, 2008

[26]A. NEKTARIOS, N.A. ASPRAGATHOS, *Optimal location of a general position and orientation end-effector's path relative to manipulator's base, considering velocity performance*, Robotics and Computer-Integrated Manufacturing, **26**, (2010) 162–173

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

A numerical approximation to the solution of the first Painlevé equation of fractional order

Vedat Suat Erturk[1]

[1] *Department of Mathematics,Ondokuz Mayis University, 55200, Samsun,Turkey*

emails: vserturk@omu.edu.tr

Abstract

This paper presents fractional differential transform method, which is a semi-analytical technique, for obtaining the numerical solution of the first Painlevé equation of fractional order.The fractional derivatives are described in the Caputo sense. The method's results are verified by calculating the residual error and explicitly reveal the efficiency and accuracy of the suggested technique.

Key words: Fractional differential transform method, Caputo fractional derivative, fractional calculus,first Painlevé equation

1. Introduction

Fractional calculus theory is a branch of the mathematical analysis that studies the possibility of taking real number powers of the differentiation and the integration operators. This generalized calculus is one of the most valuable and suitable tools to refine the description of numerous physical phenomena in science and engineering, which are indeed nonlinear. In mechanics, for example, fractional-order derivatives have been successfully used to model damping forces with memory effect or to describe state feedback controllers [1-4]. In particular, the 1/2-order derivative or 3/2-order derivative describe the frequency-dependent damping materials quite satisfactorily[5]. In fact, many physical phenomena can be modelled by fractional differential equations (FDEs), which have different applications in various areas of science and engineering such as thermal systems, turbulence, image processing, fluid flow, mechanics, and viscoelastic[1-4].
The Painlev´e equations were first formulated by Paul Painlevé [6]. It was found that, all of these equations have general solutions in terms of classical special

 ISBN: 978-84-608-6082-2

functions or elliptic functions, except for six special equations which are called the Painlev´e equations[7]. These six equations have a great variety of interesting properties and applications, for example [8] and reference therein. When general solution is impossible, one has to apply numerical techniques in order to approximate to solution. Due to this, some authors have proposed numerical techniques to approximate to the solutions of these equations.

Ellahi et al. [9] used Homotopy Analysis Method for Painlevé equation II while Behzadi [10] used various approximate analytical methods and their modifications for Painlevé equation I. Fornberg et al. [11] applied Taylor/Padé-based ODE initial value solver for Painlevé equation I. Moreover, Hesameddini et al. [12] used the Legendre wavelet method to solve Painlevé equation II while Abramov et al. [13] applied a numerical technique, which is referred as the successive elimination of singularities, to solve all the six Painlevé equations.

The first type of the Painlevé equations, Painlevé equation I, which is formulated in the following form

$$y''(x) = 6y^2(x) + x \tag{1.1}$$

with initial conditions

$$y(0) = \alpha_0, y'(0) = \alpha_1. \tag{1.2}$$

This problem arises in various physical contexts including the critical behaviour near gradient catastrophe for the focusing nonlinear Schrodinger equation[14].

Now we introduce fractional order into the integer order ordinary differential equation given in Eq. (1.1). The new equation is described by the following fractional order differential equation

$$D^{\alpha}_{x_0} y(x) = 6y^2(t) + t \tag{1.3}$$

where $D^{\alpha}_{x_0}(.)$ is the fractional derivative operator in the Caputo sense and α is a parameter describing the order of the fractional time derivative with $\alpha \in (0,2]$, subject to the same initial conditions given in Eq.(1.2). Frankly, the equation of integer-order can be viewed as a special case of the fractional-order equation by putting the fractional order of the derivative equal to two. In other words, the ultimate behavior of the fractional system response must converge to the response of the integer order version of the equation.

The major aim of this study is to analyse Eq. (1.3) and try to come to some conclusions regarding the behavior of its solutions. To the best of our knowledge, this work introduces the first available numerical solution for the first Painlevé equation of fractional-order. For this reason, we intend to obtain the approximate solution of Eq. (1.3) subject to the initial conditions (1.2) via fractional differential transform method.

This work is organized as follows. In Section 2, we present some necessary definitions and notations related to fractional calculus. In Section 3, we introduce the fractional differential transform method used in this paper to obtain the approximate semi-analytical solution for the fractional differential equations (1.3). In Section 4, the proposed method is applied to the problems (1.3)-(1.2) while

numerical simulations are presented graphically in Section 5. Finally, the conclusion is given in Section 6.

2. Preliminaries

In this section, we give some basic definitions and properties of the fractional calculus theory which are used further in this work.
There are several definitions of a fractional derivative of order $\alpha > 0$ [2]. The two most commonly used definitions are the Riemann–Liouville and Caputo. Each definition uses Riemann–Liouville fractional integration and derivatives of whole order. The difference between the two definitions is in the order of evaluation. Riemann–Liouville fractional integration of order α is defined as

$$J_{x_0}^{x} f(x) = \frac{1}{\Gamma(\alpha)} \int_{x_0}^{x} (x-t)^{\alpha-1} f(t)dt, \ \alpha > 0, \ x > 0 . \tag{2.1}$$

The next two equations define Riemann–Liouville and Caputo fractional derivatives of order α, respectively,

$$D_{x_0}^{\alpha} f(x) = \frac{d^m}{dx^m}[J^{m-\alpha} f(x)], \tag{2.2}$$

$$D_{*x_0}^{\alpha} f(x) = J^{m-\alpha}\left[\frac{d^m}{dx^m} f(x)\right], \tag{2.3}$$

where $m-1 < \alpha \leq m$ and $m \in Z^+$. For now, the Caputo fractional derivative will be denoted by D_*^{α} to maintain a clear distinction with the Riemann–Liouville fractional derivative. The Caputo fractional derivative first computes an ordinary derivative followed by a fractional integral to achieve the desired order of fractional derivative. The Riemann–Liouville fractional derivative is computed in the reverse order.We have chosen to use the Caputo fractional derivative because it allows traditional initial and boundary conditions to be included in the formulation of the problem, but for homogeneous initial condition assumption, these two operators coincide. For more details on the geometric and physical interpretation for fractional derivatives of both the Riemann–Liouville and Caputo types see [2].

3. Fractional differential transform method

In this section, we introduce the fractional differential transform method used in this paper to obtain approximate analytical solutions for Eqs. (1.3)-(1.2). This method has been developed in [15] as follows:
The fractional differentiation in Riemann–Liouville sense is defined by

 ISBN: 978-84-608-6082-2

$$D_{x_0}^{q} f(x) = \frac{1}{\Gamma(m-q)} \frac{d^m}{dx^m} [\int_{x_0}^{x} \frac{f(t)}{(x-t)^{1+q-\mathrm{m}}} dt], \tag{3.1}$$

for $m-1 < \alpha \leq m, m \in Z^{+}, x > x_0$. Let us expand the analytical and continuous function $f(x)$ in terms of a fractional power series as follows:

$$f(x) = \sum_{k=0}^{\infty} F(k)(x-\mathrm{x}_0)^{\frac{k}{\alpha}}, \tag{3.2}$$

where α is the order of fraction and $F(k)$ is the fractional differential transform of $f(x)$.

In order to avoid fractional initial and boundary conditions, we define the fractional derivative in the Caputo sense. The relation between the Riemann-Liouville operator and Caputo operator is given by

$$D_{*x_0}^{q} f(x) = D_{x_0}^{q} \left[f(x) - \sum_{k=0}^{m-1} \frac{1}{k!} (x-\mathrm{x}_0)^k f^{(k)}(x_0) \right]. \tag{3.3}$$

Since $f(x) = f(x) - \sum_{k=0}^{m-1} \frac{1}{k!} (x-\mathrm{x}_0)^k f^{(k)}(x_0)$ in Eq. (3.1) and using Eq. (3.3), we obtain fractional derivative in the Caputo sense [16] as follows:

$$D_{*x_0}^{q} f(x) = \frac{1}{\Gamma(m-q)} \frac{d^m}{dx^m} \left\{ \int_{x_0}^{x} \left[\frac{f(t) - \sum_{k=0}^{m-1} \frac{1}{k!} (t-t_0)^k f^{(k)}(x_0)}{(x-t)^{1+q-\mathrm{m}}} \right] dt \right\}. \tag{3.4}$$

Since the initial conditions are implemented to the integer order derivatives, the transformation of the initial conditions are defined as follows:

$$F(k) = \begin{cases} \text{If } k/\alpha \in Z^{+}, \dfrac{1}{(k/\alpha)!} \left[\dfrac{d^{k/\alpha} f(x)}{dx^{k/\alpha}} \right]_{x=x_0} & \text{for } k = 0,1,\ldots,(q\alpha - 1) \\ \text{If } k/\alpha \notin Z^{+}, \ 0 \end{cases} \tag{3.5}$$

where, q is the order of fractional differential equation considered. The following theorems that can be deduced from Eqs. (3.1) and (3.2) are given below, for proofs and details see [15].

Theorem 1. If $f(x) = g(x) \pm h(x)$, then $F(k) = G(k) \pm H(k)$.

Theorem 2. If $f(x) = g(x).h(x)$, then $F(k) = \sum_{\mathrm{l}=0}^{k} G(l)H(k-l)$.

Theorem 3. If $f(x) = g_1(x)g_2(x)\ldots g_{m-1}(x)g_m(x)$, then

 ISBN: 978-84-608-6082-2

$$F(k)=\sum_{k_{n-1}=0}^{k}\sum_{k_{n-2}=0}^{k_{n-1}}\ldots\sum_{k_2=0}^{k_3}\sum_{k_1=0}^{k_2}G_1(k_1)G_2(k_2-k_1)\ldots G_{n-1}(k_{n-1}-k_{n-2})G_n(k-k_{n-1}).$$

Theorem 4. If $f(x)=(x-x_0)^p$, then $F(k)=\delta(k-\alpha p)$ where,

$$\delta(k)=\begin{cases}1, & \text{if } k=m,\\ 0, & \text{if } k\neq m.\end{cases}$$

Theorem 5. If $f(x)=D_{x_0}^{q}g(x)$, then $F(k)=\dfrac{\Gamma(q+1+k/\alpha)}{\Gamma(1+k/\alpha)}G(k+\alpha q).$

4. The fractional differential transform solution of Painlevé Equation I

By taking the differential transform of Eq. (1.3), we obtain

$$Y(k+\alpha q)=\frac{\Gamma(1+k/\alpha)}{\Gamma(q+1+k/\alpha)}\left[6\sum_{l=0}^{k}Y(l)Y(k-l)+\delta(k-\alpha)\right]. \qquad (1.3)$$

5. Numerical results

In this section, as mentioned above, we use the method proposed in Ref.[15] for the numerical simulations of Eq. (1.3). This method is a very effective tool to give numerical solutions of fractional order differential equations[17-18]. It may be used both for linear and nonlinear problems encountered in science and engineering.

We take $\alpha_0=1$ and $\alpha_1=0$ as the initial conditions.

6. Conculusion

In this paper, the first Painlevé equation of fractional order is studied and its approximate solution is presented.

7. References

[1] F. MAINARDI, *Fractional Calculus and Wawes in Linear Viscoelasticity,* Imperial College Press, London, 2010.

[2] I. PODLUBNY, *Fractional Differential Equations*, Academic Press, San Diego, CA, USA, 1999.

[3] S.G. SAMKO, A.A. KILBAS, O.I. MARICHEV, *Fractional Integrals and Derivatives Theory and Applications*, Gordon and Breach, New York, 1993.

 ISBN: 978-84-608-6082-2

[4] A. KILBAS, H. SRIVASTAVA, J. TRUJILLO, *Theory and Applications of Fractional Differential Equations*, Elsevier, Amsterdam, Netherlands, 2006.
[5] R.L. BAGLEY, P.J. TORVIK, *On the appearance of the fractional derivative in the behavior of real materials*, J. Appl. Mech. **51** (1984) 294-298.
[6] P. PAINLEVé, *Mémoire sur les équations différentielles dont l'intégrale générale est uniforme*, Bull. Soc. Math. France **28** (1900), 201-261.
[7] P. A. CLARKSON, *Special polynomials associated with rational solutions of the fifth Painlevé equation*, J.Comput. Appl. Math.**178** (2005) 111-129.
[8] N. A. KUDRYASHOV, *The second Painlevé equation as a model for electric field in a semiconductor*, Phys. Lett. A, **233** (1997) 397-400.
[9] R. ELLAHI, S. ABBASBANDY, T. HAYAT, A. ZEESHAN, *On comparison of series and numerical solutions for second Painlevé equation*, Numer. Methods Partial. Differ. Equ. **26** (2010) 1070-1078.
[10] S. S. BEHZADI, *Convergence of Iterative Methods for Solving Painlevé Equation I*, Appl. Mth. Sci. **4**, 1489–1507 (2010).
[11] B.FORNBERG, J.A.C. WEIDEMAN, *A numerical methodology for the Painlevé equations*, J. Comput. Phys. **230** (2011) 5957–5973.
[12]E. HESAMEDDINI,S. SHEKARPAZ, *Wavelet solutions of the second Painlevé equation*, IJST **A4** (2011) 287-291.
[13] A. A. ABRAMOV, L.F.YUKHNO, *A method for the numerical solution of the Painlevé equations*, Comput. Math. & Math. Phys, **53** (2013) 540-563.
[14] B. DUBROVIN, T. GRAVA, and C. KLEIN, *On universality of critical behavior in the focusing nonlinear Schrodinger equation,elliptic umbilic catastrophe and the tritronqueé solution to the Painlevé-I equation*, J. Nonlinear Sci. **19** (2009) 57-94.
[15] A. ARIKOGLU, I. OZKOL, *Solution of fractional differential equations by using differential transform method*, Chaos Sol. Fract. **34** (2007) 1473-1481.
[16] M. CAPUTO, *Linear models of dissipation whose Q is almost frequency independent II*, Geophys. J. Roy. Astronom. Soc. **13** (1967) 529–539.
[17] M.MERDAN, A.GOKDOGAN, A.YILDIRIM, *On numerical solution to fractional non-linear oscillatory equations*, Meccanica, **48**(2013) 1201-1213.
[18] V.S.ERTURK,G.ZAMAN, S.MOMANI, A numeric–analytic method for approximating a giving up smoking model containing fractional derivatives, Comput.Math. Appl. 64(2012) 3065-3074.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Extension of confidence bands based on the exact distribution of the order statistics for Normal S-P Plots

María Dolores Estudillo-Martínez[1], Sonia Castillo-Gutiérrez[1] and Emilio Lozano-Aguilera[1]

[1] *Department of Statistics and Operation Research, University of Jaén*

emails: mdestudi@ujaen.es, socasti@ujaen.es, elozano@ujaen.es

Abstract

Probability plots are used as graphical techniques of goodness-of-fit. Confidence bans are included in a normal probability plot to detect non-normality in a set of observations with the advantage that the conclusion is not influenced by the subjectivity of the observer. In this contribution we have developed the extension of the confidence bands based on the exact distribution of the order statistics for Normal S-P Plots.

Key words: confidence bands, Normal Q-Q Plot, Normal S-P Plot, goodness-of-fit, graphical techniques

1. Introduction

In 1983, Michael [1] proposed a new probability plots named stabilized probability plot or S-P Plot. This graph appears as a transformation of P-P Plots with the aim of stabilizing the variance of the represented points. For example, when the theoretical distribution is the normal distribution, in Q-Q Plots, the points nearest to the centre of the graph have variances lower than the farthest. In case of P-P Plots takes place on the contrary. With this new graph, this drawback is avoided.

Michael proposed acceptance regions for Normal S-P Plots. In this work, we perform the extension of confidence bands for Normal S-P Plots, through the analysis of variability in Normal Q-Q Plot based on the exact distribution of the order statistics.

 ISBN: 978-84-608-6082-2

2. Normal S-P Plot

Given a set of ordered observations, $x_{(1)},...,x_{(n)}$ and $\Phi(x)$ the normal distribution function, the steps to construct a Normal S-P Plot are the following:

1) Determine the value of the plotting positions (p_i). In this step we have to choose one definition for p_i; we choose Hazen's definition [2], i.e.:

$$p_i = \frac{i-0.5}{n} \qquad i = 1,...,n$$

2) Calculate abscissa values by applying the transformation:

$$r_i = \left(\frac{2}{\pi}\right) arcsen\left(p_i^{1/2}\right) \qquad i = 1,...,n$$

3) Determine:

$$v_i = \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right) \qquad i = 1,...,n$$

Using the corresponding unbiased estimators for the parameters μ and σ of the normal distribution.

4) Calculate the y-axis values through the application of the transformation:

$$s_i = \left(\frac{2}{\pi}\right) arcsen\left(v_i^{1/2}\right) \qquad i = 1,...,n$$

5) Draw, for each *i=1,...,n*, the point:

$$\left(r_i, s_i\right)$$

3. Extension of confidence bands based on the exact distribution of the order statistics for Normal S-P Plots

In 1983, Michael proposed acceptance regions for Normal S-P Plots. In this section we develop the extension of the confidence bands in Normal Q-Q Plots based on the exact distribution of the order statistics [3], to be applied to Normal S-P Plots.

Given a set of ordered observations, $x_{(1)},...,x_{(n)}$ and $\Phi(x)$ the normal distribution function, the steps to construct confidence bands in a Normal S-P Plot based on the exact distribution of the order statistics are as follows:

1) Fix a significance level α.
2) Draw a Normal S-P Plot for the observations using the corresponding unbiased estimators for the parameters μ and σ of the normal distribution, following the steps in section 2.

 ISBN: 978-84-608-6082-2

3) Determine, for each i $(1 \le i \le n)$, the values $p_1^{(i)}(\alpha)$ and $p_2^{(i)}(\alpha)$, which are the quantiles of order $(\alpha/2)$ and $1-(\alpha/2)$, respectively, of a Beta distribution $Beta(i, n-i+1)$.
4) Apply the transformation proposed by Michael to stabilize the variance in the values obtained in the previous step:

$$l_i = \frac{2}{\pi} arcsen\left(\sqrt{p_1^{(i)}(\alpha)}\right) \qquad i = 1,...,n$$

$$u_i = \frac{2}{\pi} arcsen\left(\sqrt{p_2^{(i)}(\alpha)}\right) \qquad i = 1,...,n$$

5) For each *i,* draw the following lower and upper limits through the points (r_i, l_i) and (r_i, u_i).
6) Join those points to create the effect of confidence bands.

The application of this graph with the confidence bands is that we will reject the hypothesis of normality if at least $\alpha\%$ of the observations fall outside the confidence bands.

4. References

[1] J.R. MICHAEL, *The Stabilized Probability Plot*, Biometrika **70(1)** (1983) 11-17.
[2] A. HAZEN, *Flood Flows. A Study of Frequencies and Magnitudes,* Wiley, New York, 1930.
[3] M.D. ESTUDILLO-MARTÍNEZ, S. CASTILLO-GUTIÉRREZ AND E.D. LOZANO-AGUILERA, *New confidence bands in Q-Q Plots to detect non-normality,* Int. J. Comput. Math. **90(10)** (2013) 2137-2146.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Network model for simulating 1-D soil consolidation processes under load-unload conditions

García G.[1], Alhama I.[1] and Sánchez J.F.[2]

[1] *Civil Engineering Department, ETSIC*
Technical University of Cartagena
[2] *Applied Physics Department, ETSIA*
Technical University of Cartagena

emails: gonzalo.garcia@upct.es , ivan.alhama@upct.es , juanfsanchez@upct.es

Abstract

Making use of the powerful mathematical algorithms implemented in the modern circuit simulation codes and following the steps of the network method, a model is designed for simulating load-unload soil consolidation process in 1-D rectangular geometries. The model is run in the simulation free code Ngspice. The change from load to unload condition is carried out by switches, which set the value of consolidation coefficient of each state. Local and average degree of consolidation as well as soil settlements are given at any time, which provides mechanical soil engineers an interesting tool for the design of consolidation processes. An application is shown to test the reliability of the model and show how the switching time that turns the load to unload condition strongly influences the total time to get a nearly constant settlement.

Key words: soil consolidation, numerical solution, network method, load-unload processes

1. Introduction

The soil consolidation process in lineal 1-D problems is an old subject, well studied in most books of soil mechanics which provide analytical solutions in many cases [1,2]. In addition, when applied to load (followed by) unload conditions these analytical solutions are not easy to manage by practical engineers. In this

 ISBN: 978-84-608-6082-2

communication, we present a numerical solution of this kind of processes, easy to implement by the user. This model, based on network simulation method [3], is capable of solving these scenarios with enough reliability and negligible computational times, using a free code of circuit simulation.

Network simulation method is a numerical tool that has been broadly used in last years for the solutions of many other lineal and non-lineal, coupled and un-coupled engineering processes, such as in heat transfer [4], tribology [5] and elasticity [6]. The models are designed starting from the finite-difference differential equations that derive from the spatial discretization of the governing equations that set the mathematical model. After stablishing the analogy between the variables of the physical process and those of the network model, each term of the equations is assumed as an electric current that balances with the currents of the other terms in a common node of the volume element, whose voltage provides the solution of the problem. The model, formed by as many circuits as equations – or dependent variables – are in the mathematical model, is generally written as a text file following a few programing rules since the expressions of the terms of the equations are quite similar from one process to other. This requires a few electrical devices to implement the complete model. Once this is designed, it is run in a suitable code of circuit simulation making use of the powerful numerical algorithms integrated in this kind of software.

2. The physical and mathematical model

The governing equation assumes the hypothesis of Terzaghi [1]. For 1-D geometry, Figure 1 (z=0 at the bottom of the domain, H is the total thickness), we will assume that the water flow leaves the domain through the soil surface (top boundary), while the bottom boundary is impermeable. The initial load q_o (N/m^2) is applied at the surface along t_o years, after which it is totally or partially withdrawn. Doing that the removed load (q_r) be a fraction of the initial total load (q_o), we can write $q_r = C \cdot q_o$, being C a coefficient between 0 and 1.

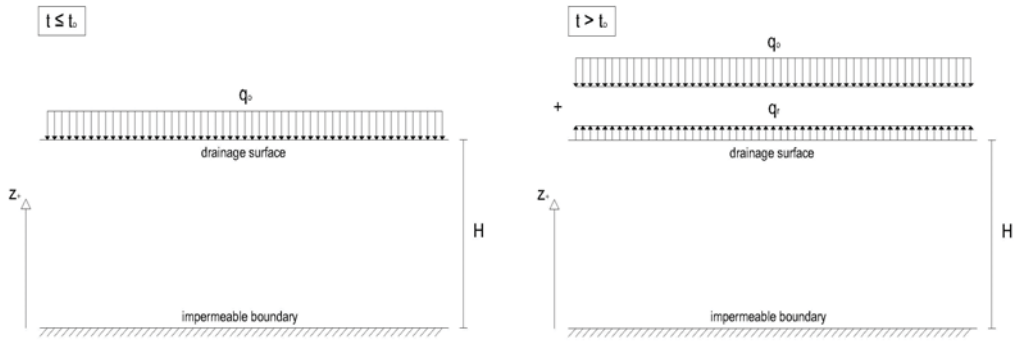


Figure 1. Physical scheme of the load-unload problem

 ISBN: 978-84-608-6082-2

The excess pore pressure of the soil at time t_{o-} (just immediately before the unload), $u(z,t_{o-})$, for which the load condition finishes, is not uniform. At t_{o+}, just when the unload period starts, the initial condition of the soil is $u(z,t>t_o) = u(z,t_{o-}) - q_r$. The changing value of $u(z,t>t_o)$ causes changing gradients in the whole domain that force the water to move from one to adjacent pores; these gradients may change their sign with time but, finally, tends to zero (as well as u) asymptotically. So that, once the unload starts and time increases, if for a given depth the gradient of the excess pore pressure is positive the soil will tend to swell, doing that the water returns to the soil pores. On the contrary, if the gradient remains negative the consolidation process continues.

Under the above hypothesis, the governing equations and boundary and initial conditions express in the form

$$\frac{\partial u}{\partial t} = c_{v1}\left(\frac{\partial^2 u}{\partial z^2}\right) \qquad \text{(Consolidation process)} \qquad (1a)$$

$$\frac{\partial u}{\partial t} = c_{v2}\left(\frac{\partial^2 u}{\partial z^2}\right) \qquad \text{(Swelling process)} \qquad (1b)$$

$$u_{(z=H,t\le t_0)} = 0, \qquad \frac{\partial u}{\partial z}_{(z=0)} = 0, \qquad u_{(z,t=0)} = q_o \qquad (2a)$$

$$u_{(z=H,t>t_0)} = 0, \qquad \frac{\partial u}{\partial z}_{(z=0)} = 0, \qquad u_{(z,t=t_{o+})} = u_{(z,t=t_o)} - q_r \qquad (2b)$$

where c_{v1} and c_{v2} are the vertical and swelling soil consolidation coefficients, respectively, defined as

$$c_{v1} = \frac{k_v\,(1+e_o)}{\gamma_w\, a_{v1}} \qquad c_{v2} = \frac{k_v\,(1+e_o)}{\gamma_w\, a_{v2}}$$

where k_v is the constant permeability of the soil, e_o the initial void ratio, γ_w the density of the water and a_{v1} and a_{v2} the compressibility and swelling coefficients, respectively.

As mentioned, the process of swelling does not start at all points of the medium at t_o. For the case of a complete withdrawal of the load ($q_r = q_o$), the gradient of the excess pore pressure remains negative at the domain except near the surface at the beginning of the unload period. So that the soil continues its consolidation nearly at all points. Nevertheless, paying attention to the closest region to the surface for $t>t_o$, we see that a change in the sign of u (and its gradient) gives rise immediately, forcing the soil to swell at that subdomain. As time continues increasing, this swelling region of the soil extends to cells immediately below until, finally, the whole package of soil is swelling, Figures 2 and 3.

 ISBN: 978-84-608-6082-2

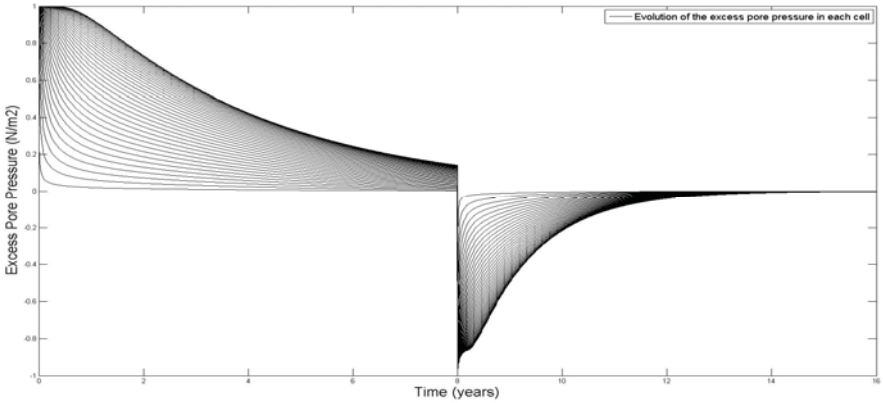


Figure 2. Negative excess pore pressure after load withdrawal.

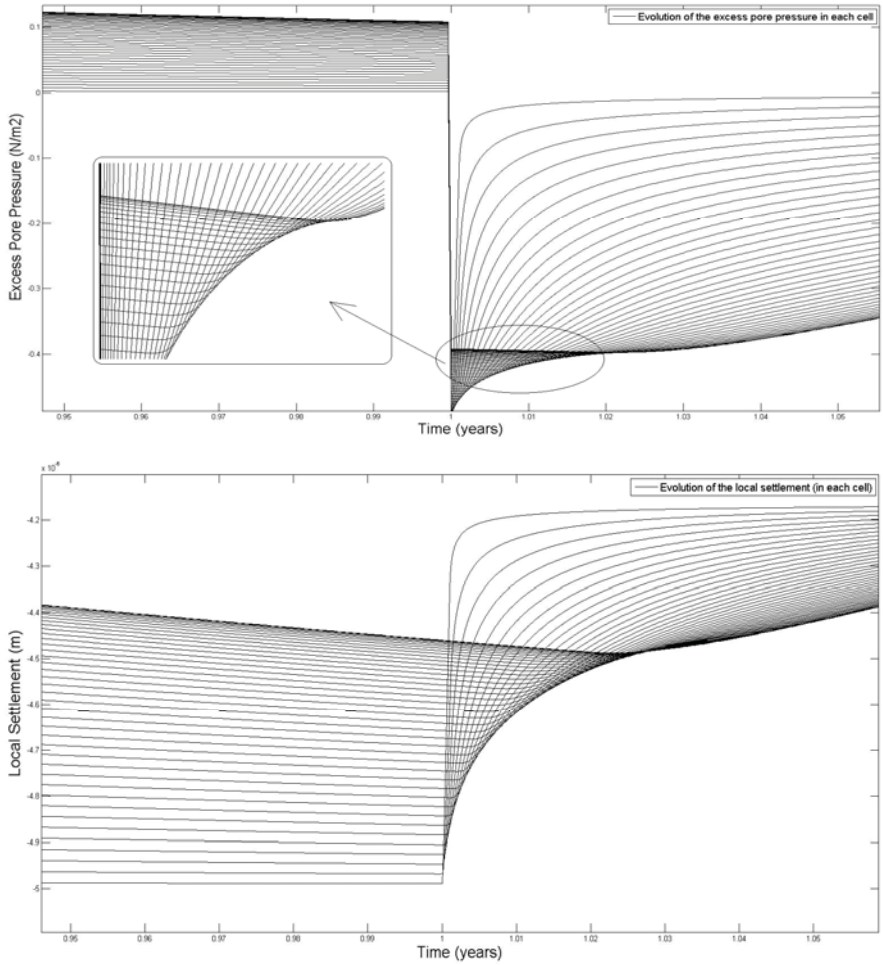


Figure 3. Local excess pore pressure and settlement before and after unload

 ISBN: 978-84-608-6082-2

3. The design of the network model

Equation (1a), in terms of the spatial derivative components, has the form

$$\frac{\partial u}{\partial t} = c_{v1}\,\frac{\partial}{\partial z}\left\{\left[\frac{\partial u}{\partial z}\right]_{z^+} - \left[\frac{\partial u}{\partial z}\right]_{z^-}\right\} \tag{3}$$

with z^+ and z^- being the entering and leaving boundaries of the cell. In finite-differences, with the nomenclature of Figure 4, this equations can be written as

$$\frac{\partial u}{\partial t} = \left[\frac{u_{i+\Delta z/2} - u_i}{\frac{(\Delta z)^2}{2c_{v1}}}\right] - \left[\frac{u_i - u_{i-\Delta z/2}}{\frac{(\Delta z)^2}{2c_{v1}}}\right] \tag{4}$$

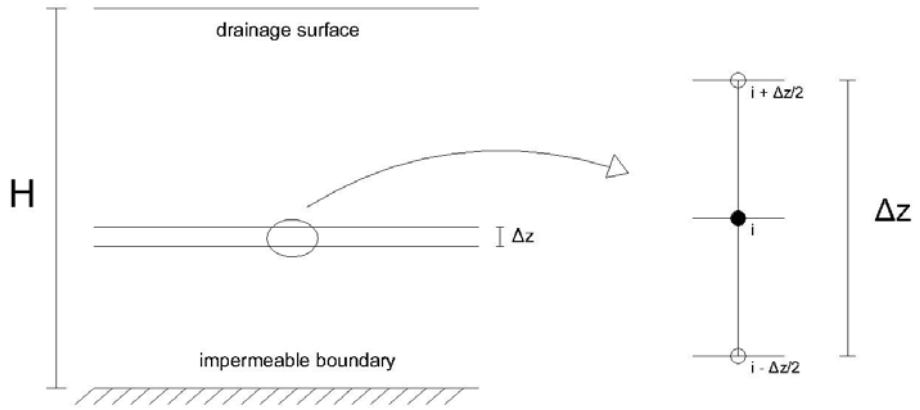


Figure 4. Cell nomenclature used for this problem

In the electric analogy, each term of this equation defined as

$$i_C = \frac{\partial u}{\partial t}, \qquad i_{R+\Delta z_1} = \frac{u_{i+\Delta z/2} - u_i}{\frac{(\Delta z)^2}{2c_{v1}}}, \qquad i_{R-\Delta z_1} = \frac{u_i - u_{i-\Delta z/2}}{\frac{(\Delta z)^2}{2c_{v1}}}$$

is an electric current that converges with the others at a common node of the cell so that the balance $i_C = i_{R+\Delta z_1} - i_{R-\Delta z_1}$ is satisfied. Following the theory of the network method, time derivative term, $i_C = \partial u/\partial t$, is implemented by a capacitor of capacitance unity while the other terms by resistors of value

$$R_{i+\Delta z_1} = R_{i-\Delta z_1} = \frac{(\Delta z)^2}{2c_{v1}} \qquad \text{(Load period)}$$

$$R_{i+\Delta z_2} = R_{i-\Delta z_2} = \frac{(\Delta z)^2}{2c_{v2}} \qquad \text{(Unload period)}$$

 ISBN: 978-84-608-6082-2

Coupling between cells is carried out by ideal electric contact while boundary conditions are implemented by voltage generator of zero value at the top boundary (Dirichlet condition), and by resistors of very high value at the bottom (homogeneous Neumann condition). Finally, initial condition is established by setting the voltage of capacitors to q_o.

The change from the load to the unload condition is implemented by switches that provide suitable values to the resistors. Besides, since the change from consolidation to swelling is not simultaneous in the whole domain, but delayed with depth, the action of the switches is programmed according to the sign of the current (or the u gradient), Figure 5.

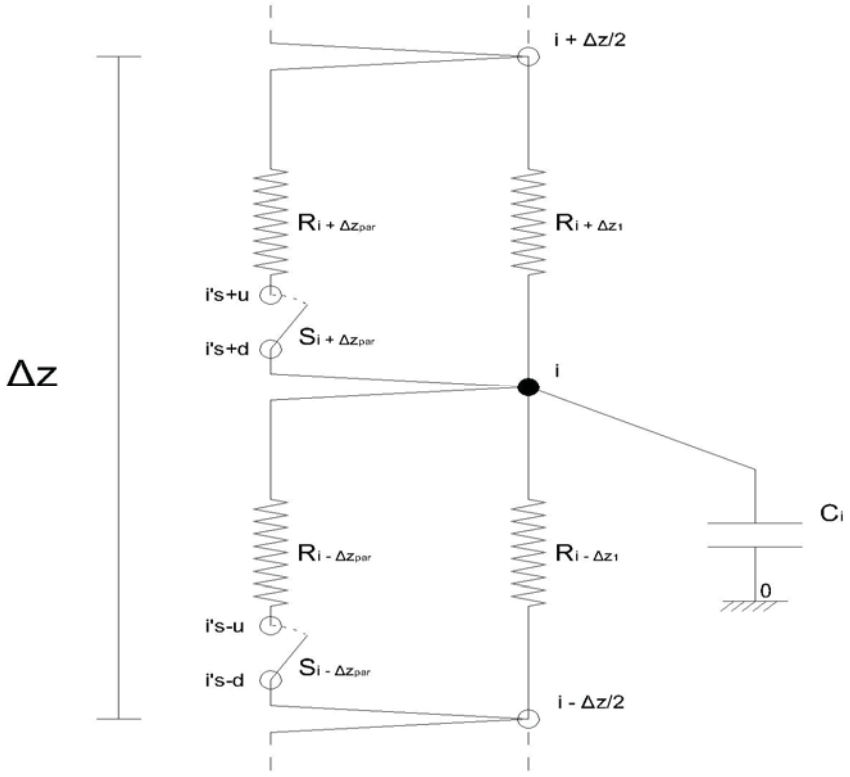


Figure 5. Network model of the cell

The complete network model is solved by the free software Ngspice [7] which provides the values of u(z,t) at any point of the model. Soil settlement is evaluated by the expression

$$S = \frac{k_v}{\gamma_w} \int_0^H \int_0^t \frac{\partial u}{\partial z} \, dt \, dz \tag{5}$$

 ISBN: 978-84-608-6082-2

4. Applications

It has been consider a real scenario of consolidation for which a uniform load of 20 kN/m^2 is applied over a clay layer of 3 m thick, supported by a rock material which is assumed rigid and impermeable. The properties of this clay are: $a_{v1} = 2e^{-3}$ m^2/kN, $a_{v2} = 6.66e^{-4}$ m^2/kN, $e_0 = 1$, $c_{v1} = 1$ m^2/year y $c_{v2} = 3$ m^2/year.

The expression of the final settlement S (for a constant load) is given by equation (6) and has a value of 0.06 m.

$$S = H\, q_o\, \frac{a_{v1}}{1+e_0} \tag{6}$$

Using the semi-analytical solutions for the average degree of consolidation, whose results are confirmed by simulations carried out, the final settlement would be reached after 20 years, while 90% of this settlement is reached at 7.5 years, still having a 6 mm settlement to develop from this moment. Figure 6.

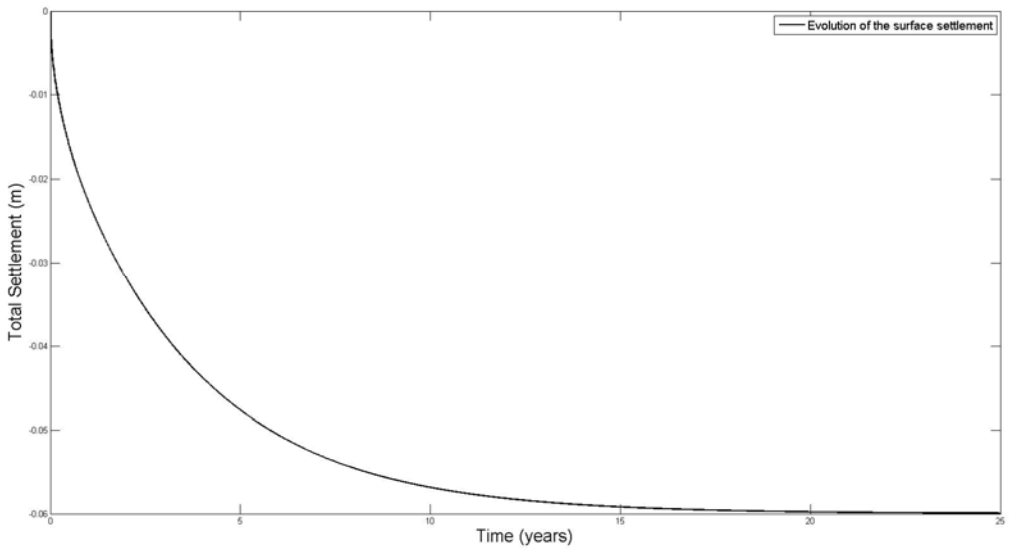


Figure 6. Settlement evolution. H=3m, q_o=20kN/m^2, c_{v1}=1 m^2/year, a_{v1}=2e^{-3} m^2/kN

In order to accelerate the consolidation process to achieve the looked for settlement in a lower time, an initial preload of 60 kN/m^2 is applied and, at a certain time, 40 kN/m^2 are withdrawn so that the resultant applied load from that moment are the 20 kN/m^2 initially projected.

Figures 7 to 9 show the results of the simulations made for the load withdrawal times of 0.5 years (t_o), 1.5 years and 2 years, respectively. For the first case, the final settlement (0.075 m) is achieved after 15 years, while a fraction of 90% of this

 ISBN: 978-84-608-6082-2

value takes place in 6 years (still having a 7.5 mm settlement to develop). For t_o=1.5 years, the final settlement (0.082 m) is reached after 10 years, while 90% of this value takes place in 1.2 years. For this same case, 95% of settlement happens only in 1.9 years, still having 4 mm to settle from this moment. Finally, for the case of load withdrawal after 2 years, the final settlement happens after 5 years (0.087 m). However, in this case the final settlement is reached as a result of a swelling process in the soil layer. In comparison with the former cases, the preload application has lasted such a time that it has caused a total settlement higher than the settlement that will be reached once the preload is withdrawn. However, for the first two cases (early withdrawal), we find that the soil begins to swell, whereas after some time the soil returns to consolidate.

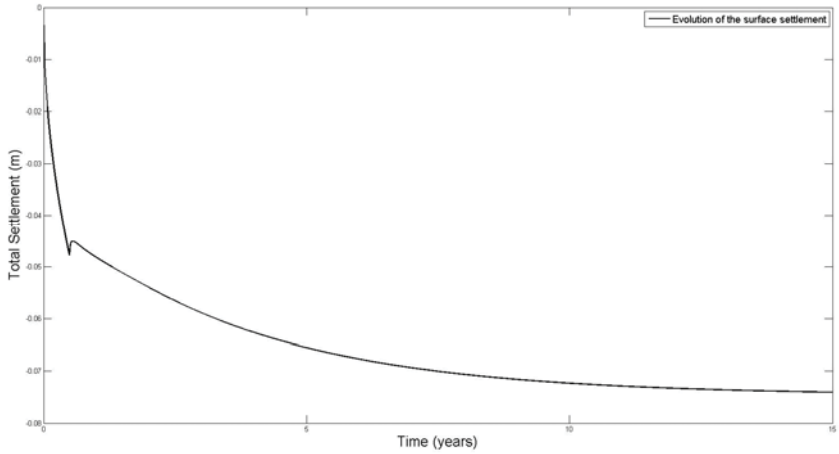


Figure 7. Settlement evolution. H=3m, q_o=60kN/m^2, c_{v1}=1 m^2/year, a_{v1}=2e^{-3} m^2/kN, q_r=40kN/m^2, c_{v2}=3 m^2/year, a_{v2}=6.66e^{-4} m^2/kN, t_o=0.5 years.

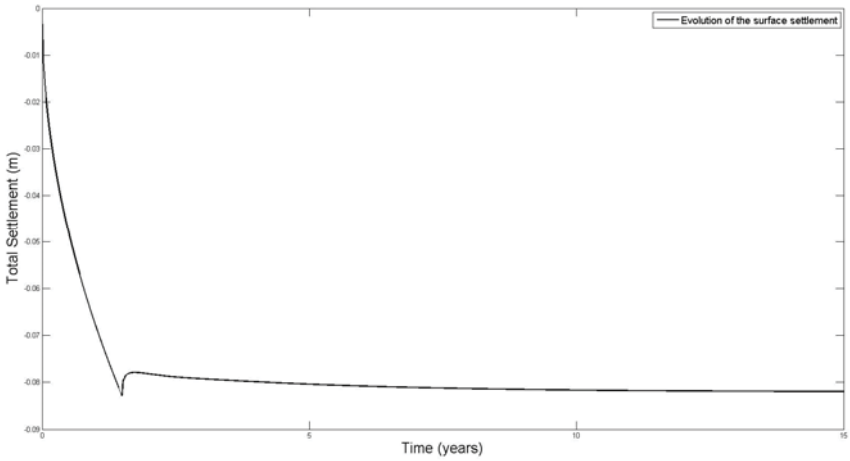


Figure 8. Settlement evolution. H=3m, q_o=60kN/m^2, c_{v1}=1 m^2/year, a_{v1}=2e^{-3} m^2/kN, q_r=40kN/m^2, c_{v2}=3 m^2/year, a_{v2}=6.66e^{-4} m^2/kN, t_o=1.5 years.

 ISBN: 978-84-608-6082-2

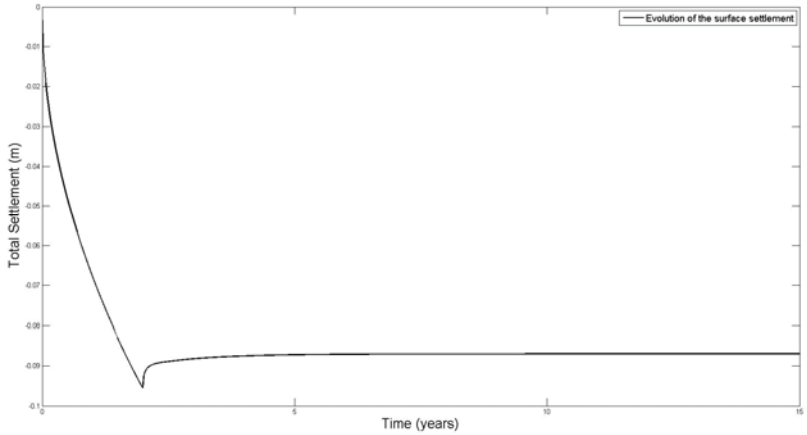


Figure 9. Settlement evolution. H=3m, q_o=60kN/m^2, c_{v1}=1 m^2/year, a_{v1}=2e^{-3} m^2/kN, q_r=40kN/m^2, c_{v2}=3 m^2/year, a_{v2}=6.66e^{-4} m^2/kN, t_o=2 years.

5. Final comments and conclusions

A good design of load-unload consolidation processes, choosing suitably both the values of the loads and the time in which the process change from one condition to another, allows the engineer to reach the desired settlements in a considerable lower time saving costs in the construction. The network method provides a reliable and powerful model to simulate this kind of problems with no other restrictions. The existence of consolidation and swelling processes simultaneously in the domain and the continuous change of this condition as time progress makes difficult to analyse this scenario. The model is tested by an application where all the phenomena involved as well as the final settlements of different design parameters are solved.

6. References

[1] K. Terzaghi. *Theoretical Soil Mechanics*. John Wiley and Sons, 1943.
[2] P.L. Berry and D. Reid. *An Introduction to Soil Mechanics*. McGraw-Hill, London, 1988.
[3] C.F. González Fernández. *Network simulation method*, Ed. J. Horno, Research Signpost Trivandrum (2002).
[4] M. Cánovas et al. *Numerical simulation of Nusselt-Rayleigh correlation in Bénard cells. A solution based on the network simulation method.* International Journal of Numerical Methods for Heat & Fluid Flow **25-5** (2015) 986-997.
[5] F. Marín et al. *Modelling of nanoscale friction using network simulation method***,** CMC: Computers, Materials & Continua **43** (2014) 1-19.
[6] J.L. Morales et al. *Numerical solutions of 2-D linear elastostatic problems by network method*, Computer modeling in Eng. & Sci. **76-1** (2011) 1-18.
[7] Ngspice, (2013): mixed-level/mixed-signal circuit simulator.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

On the Capabilities of the Open-Source HEVC Codecs

David Garcia-Lucas, Gabriel Cebrián-Márquez and Pedro Cuenca

Albacete Research Institute of Informatics (I3A), University of Castilla-La Mancha, Albacete, Spain

emails: David.Garcia72@alu.uclm.es, Gabriel.Cebrian@uclm.es, Pedro.Cuenca@uclm.es

Abstract

High Efficiency Video Coding (HEVC) was developed by the Joint Collaborative Team on Video Coding (JCT-VC) to replace the current H.264/Advanced Video Coding (AVC) standard, which has dominated digital video services in all segments of the domestic and professional markets for over ten years. In terms of rate-distortion (R-D) performance, HEVC roughly doubles the R-D compression performance of H.264/AVC but at a cost of extremely high computational and storage complexities during encoding. Since its standardization, several open-source HEVC video codecs have been developed. This paper presents a rate-distortion/complexity analysis of the open-source HEVC codecs using objective measures of assessment in order to analyse their real capabilities. The comparison was done using settings provided by the developers of each codec. Experimental results show that the HEVC codecs analysed cannot achieve an acceptable good trade-off between coding efficiency and complexity.

Key words: HEVC codecs, Evaluation, Computational Cost

1. Introduction

In the last years, H.264/*Advanced Video Coding* (AVC) [1] has been the most widespread video compression standard for all types of applications and scenarios. Nevertheless, the continuous market demands for a better quality of experience and the advent of new video formats such as the *Ultra High Definition* (UHD) resolution motivated the development of the *High Efficiency Video Coding* (HEVC) standard [2]. Established by the *Joint Collaborative Team on*

 ISBN: 978-84-608-6082-2

Video Coding (JCT-VC) in early 2013, HEVC roughly doubles the *Rate Distortion* (R-D) performance of H.264/AVC. This improvement in coding efficiency comes, however, at the expense of an extremely high computational complexity [3]. The popularity of those emergent formats requires of higher compression efficiency, and demands higher performance computation. However, some constraints to the video processor devices have to be considered, such as the reduced battery capacity of the hand-held and mobile devices.

With the aim of reducing this high computational complexity, several suboptimal fast HEVC video codecs have been developed by the industry. Up to this date, several HEVC encoders have been released, but most of them are commercial products whose features and operating principles are kept confidential. Therefore, only open-source encoders are considered in this paper.

Among the existing open-source HEVC encoders, *HEVC Test Model* (HM) [4] as an HEVC reference codec is able to achieve the best coding efficiency among the existing HEVC encoders, but its object-based C++ implementation results in poor performance in terms of computational complexity. Hence, it is targeted for research and conformance testing rather than practical encoding. The commercially funded x265 [5] is the most well-known practical open-source HEVC encoder. It is based on HM C++ source code which has been enhanced by extensive assembly optimizations, multithreading, and techniques from the open-source x264 encoder. F265 [6] is another industrial HEVC encoder. It is implemented in C with assembly optimizations. Although the source code for these two commercially led projects are under open-source licenses, contributors to these projects must sign an agreement giving the companies copyright to their work. Requiring such agreements leaves room for non-commercial projects, such as Kvazaar [7], that do not require signing separate agreements to participate. Kvazaar is an academic open-source HEVC encoder initiated and coordinated by [8]. It is licensed under GNU GPLv2 license. Kvazaar uses a reverse design approach compared with x265: it has been developed from scratch using HM primarily as a reference for its encoding scheme and individual algorithm implementations. In addition, Kvazaar is developed in C. This more hardware-oriented approach eases source code acceleration, portability, and parallelization. *HEVC Open MPEG Encoder* (HomerHEVC) [9] is an open-source, real-time, multiplatform HEVC video encoder under LGPL license. Finally, DivX also presents its DivX HEVC encoder known as Divx265 [10].

This paper focuses on a comparative evaluation of the quality/computational cost of the open-source HEVC codecs using objective measures of assessment in order to analyse their real capabilities. The comparison was done using settings provided by the developers of each codec.

 ISBN: 978-84-608-6082-2

The remainder of this paper is organized as follows. Section 2 includes some technical background to the new HEVC standard. Section 3 presents the open-source HEVC video codecs under study, and then the experimental results are given in Section 4. Finally, Section 5 concludes the paper.

2. Technical Background

HEVC can be considered an evolution of the current H.264/AVC, since it maintains the same block-based hybrid approach used in all previous video compression standards. In addition, new tools have been introduced in HEVC that increase its coding efficiency compared with H.264/AVC [11].

One of the most important changes affects picture partitioning [12]. HEVC defines a new flexible *Coding Tree Unit* (CTU) structure which is a replacement of the *Macroblocks* (MBs), 16×16 pixel blocks, used in the previous standards. With the aim of achieving an optimal adaptation to the content details, CTUs can vary from a size of 64×64 pixels, to something much smaller, as it can be iteratively partitioned into four square sub-blocks of half resolution, named *Coding Units* (CUs), with a minimum allowable size of 8×8 pixels. Therefore, a CTU can be further partitioned into four depth levels, from *d=0* for 64×64 CU to *d=3* for 8×8 CUs, having 4^d CUs in each depth level. Thus, a CU in depth level *d* can be denoted as $CU_{d,k}$ (k=0,1,.., 4^d-1), and the four sub-CUs pending on $CU_{d,k}$ are denoted as $CU_{d+1,4k+i}$ (i=0,1,.., 3). In a CTU of 64×64, it can be observed that the maximum number of available CUs is $\sum_{d=0}^{d=3} 4^d$.

HEVC increases even more the flexibility of the CTU by defining two tree structures containing new unit types: the *Prediction Units* (PUs), and the *Transform Units* (TUs). For intra-picture prediction, a PU uses the same 2N×2N size as for the $CU_{d,k}$ to which it belongs, allowing it to be split into four N×N PUs only for CUs at the minimum depth level. Therefore, the PU size can range from 64×64 to 4×4 pixels. For inter-picture prediction, several non-square rectangular block shapes are available in addition to square ones, allowing eight different PU sizes (2N×2N, 2N×N, N×2N, N×N, 2N×nU, 2N×nD, nL×2N, nR×2N). The prediction residue obtained in each of the PUs is transformed using various TU sizes from 32×32 to 4×4. In Figure 1, an example of the partitioning is shown, depicting how a CTU is structured in a hierarchical tree where each CU branch ends in a leaf ($CU_{d,k}$) that is the root for the two new prediction and transform trees. Figure 2 shows the partitioning of CTU into CUs (white), PUs (green) and TUs (black) applied to the *Basketball Pass* sequence.

It should be noted that a CTU can be split into 341 different PUs ($\sum_{d=0}^{d=3} 4^d$), and each of these available PUs has to be evaluated for all intra/inter prediction modes available, and each of the obtained residual blocks can be transformed into up to three TU sizes. HEVC checks most of the PUs (inter and intra modes) to decide

 ISBN: 978-84-608-6082-2

whether it should split a CU or not by choosing the best R-D case. Furthermore, in the case of inter prediction, for each of these PU partitions a motion estimation algorithm is called. This wide range of possibilities makes HEVC much more computationally expensive than its predecessor, H.264/AVC. HEVC introduces changes in other modules too, such as intra prediction (where a total of 35 different coding modes can be selected), new image filters or new transform sizes [11], among others.

The above analysis evidences the need to reduce the *Rate-Distortion Optimization* (RDO) complexity for the HEVC intra/inter prediction, in order to make real time HEVC video codecs with the best possible performance. With the aim of reducing this huge RDO complexity, several suboptimal fast HEVC video codecs have been developed by the industry using a reduced set of prediction modes that are previously selected as candidates, in a low complexity evaluation process.

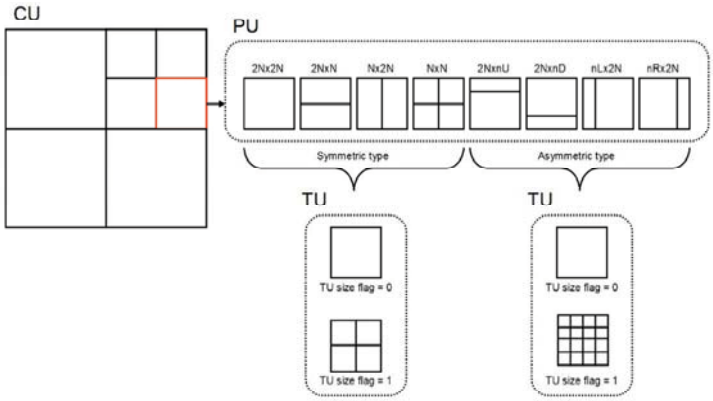


Figure 1. Partitioning of CTU into CUs, PUs and TUs.

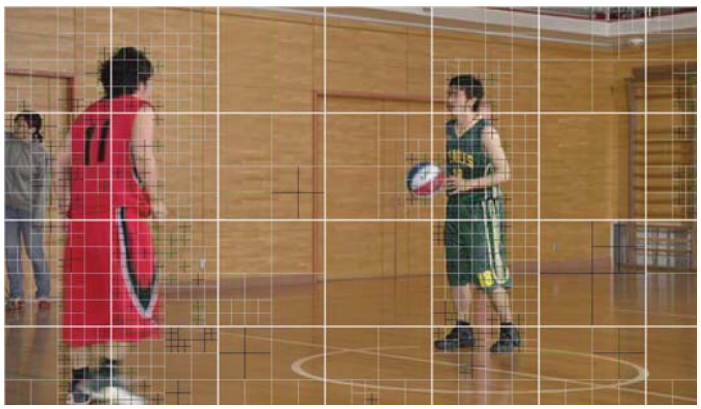


Figure 2. Partitioning of CTU into CUs (white), PUs (green) and TUs (black) applied to *Basketball Pass* sequence.

 ISBN: 978-84-608-6082-2

3. Open-Source HEVC Video Codecs

With the aim of reducing this high computational complexity, several suboptimal (from the coding efficiency point of view) open-source HEVC encoders have been developed by the industry.

Among the existing open-source HEVC encoders, **HM** [4] as an HEVC reference codec is able to achieve the best coding efficiency among the existing HEVC encoders. Recommendation ITU-T H.265.2 contains an accompanying reference software for Rec. ITU-T H.265 | ISO/IEC 23008-2 High Efficiency Video Coding [2]. The HM software includes both encoder and decoder functionality. Reference software is useful in aiding users of a video coding standard to establish and test conformance and interoperability, and to educate users and demonstrate the capabilities of the standard. For these purposes, the accompanying software is provided as an aid for the study and implementation of Rec. ITU T H.265 | ISO/IEC 23008-2. However, its object-based C++ implementation results in poor performance in terms of computational complexity. Hence, it is targeted for research and conformance testing rather than practical encoding.

x265 [5] is a commercially funded open-source implementation of the H.265/HEVC compression standard. The x265 project is led by MulticoreWare, a leading provider of high performance video software libraries. The x265 codec is available under the GNU GPL v2 license. Development is funded through license fees paid by commercial companies, who receive a commercial license which is less restrictive than the GPL v2 (with respect to open-source requirements for the system that integrates x265). The x265 team has licensed the rights to port and adapt x264 for use in x265. Most of the video encoding algorithms developed for x264 have been incorporated in x265, including rate control functions, the look-ahead function, macroblock tree, b-pyramid and adaptive quantization. It is based on HM C++ source code which has been enhanced by extensive assembly optimizations and multithreading techniques.

Vantrix, a global provider of high-performance media delivery solutions, recently announced the creation of **F265** [6] open-source project to accelerate the development of a free H.265/HEVC video encoder. Vantrix has made the source code for its HEVC codec available to the F265 project under the OSI BSD license terms. That means that whether you are an academic researcher or work for a commercial entity, you will have access to source code and the ability to freely distribute any derived works. The standard behaviour for F265 is to encode the source video at average size and quality, meaning that those looking for custom values need to manually tamper the encoding parameters. F265 also allows users to disable the stats output that can be used for profiling and benchmarking, as well as deactivate the assembly support even it is available. It is implemented in C with assembly optimizations.

 ISBN: 978-84-608-6082-2

Kvazaar [7] is an academic open-source HEVC encoder initiated and coordinated by [8] with the following primary goals: 1) coding efficiency close to HM, 2) modular structure to ease its parallelization and portability, and 3) excellent source code readability and documentation. To obtain these objectives with the highest possible encoding speed and with minimal computation/memory resources, Kvazaar is developed from scratch in C. This unique approach uses the object-based C++ implementation of HM as a reference for its encoding scheme and individual algorithm implementations, but it adopts completely new data and function call tree structures. The supported platforms are x86, x64, and PowerPC on Windows, Linux, and Mac.

HomerHEVC [9] is an open-source, real-time, multiplatform H.265/HEVC video encoder under LGPL license. The current features in its version 2.0 are: multiplatform (Linux, Windows), 8 bit-depth, intra and baseline profile, multiple references for IPP, all intra prediction modes, 2N×2N and N×N inter prediction modes, all prediction sizes (64, 32, 16, 8, 4), all transform sizes (32, 16, 8, 4), half pixel and quarter pixel precision motion estimation, rate control modes (fixed QP, CBR and VBR), deblocking filter, *Sample Adaptative Offset* (SAO), WPP and frame based parallelization for massive parallelism, sign hiding bit enabled, intra RDO and intra-inter fast R-D. Current SSE42 optimizations (intrinsics) are: Intra prediction, motion estimation, inter prediction and biprediction with 1/8 pixel chroma precision, intra prediction, reconstruction, SAD, transforms, quantization and SAO.

Finally, **DivX** and NeuLion solutions are powering the professional creation, secure distribution, and multi-screen playback of high-quality video for leading companies around the world [10]. DivX securely distributes video to multiple screens enabling over-the-top service operators and platforms to differentiate their service, reduce operating costs and extend audience reach, with support for iOS, Android, web, PC, Mac, game consoles, set-top boxes and smart TVs. DivX enables Pay TV Operators, Broadcasters, Channels, OTT Services and In-Flight Entertainment system vendors to create, securely distribute and play over-the-top video across a wide array of screens including mobile and smart TVs. The DivX Certified devices include an addressable device base of over 1 billion.

4. Rate-Distortion/Complexity Analysis

4.1. Encoding Parameters

This section aims to evaluate the rate-distortion/complexity capabilities of the open-source HEVC codecs presented in this paper. The parameters used for each encoder are listed in Table 1. Our experiments rely on the default configuration of HM 16.6 which is used as an anchor for the obtained results. With regard to the

 ISBN: 978-84-608-6082-2

rest of the open-source HEVC codecs under study, the default medium encoding presets have been used, which define the sets of parameters that achieve a trade-off between performance and coding efficiency. For a fair comparison, the experiments are conducted with a single-threaded implementation of all of them by disabling its parallelization options.

The hardware platform used in the experiments is composed of an Intel® Xeon® E5-2630L v3 CPU running at 1.80 GHz and 16 GB of main memory. The encoders have been compiled with GCC 4.8.5-4 and executed on CentOS 7 (Linux 3.10.0-327). Turbo Boost has been disabled to achieve the reproducibility of the results.

Table 1: Encoder Parameters used for Simulation.

Encoder	Parameters
HM	-c (encoder configuration file) -c (sequence configuration file) --IntraPeriod=(intra period) --QP (qp)
HomerHEVC	-widthxheight (resolution) -n_frames (frames) -frame_rate (fps) -bitrate_mode 0 -qp (qp) -gop_size (0 for AI, 1 for LP, 2 for RA) -intra_period (intra period) -n_enc_engines 1 -n_wpp_threads 1 -performance_mode (preset)
Kvazaar	--input-res (resolution) -n (frames) --input-fps (fps) --qp (qp) -p (intra period) --gop (8 for RA, lp-g4d4r4t1 for LP) --threads 1 --wpp --preset (preset)
x265	--input-res (resolution) -f (frames) --fps (fps) --qp (qp) --keyint (intra period) --bframes (0 for LP) --pools 1 --frame-threads 1 –p (preset)
Divx265	-s (resolution) -n (frames) -fps (fps) -I 1 -qp (qp) -aqo (medium)
F265	-w (resolution) -c (frames) -p "fps=(fps,1) qp=(qp) wpp=1 key-frame-spacing=(intra period) bframes=(7 for RA, 0 for LP) ref=(3 for RA, 4 for LP)"

In order to ensure a common framework for the simulations, the experiments were conducted under the *Common Test Conditions and Software Reference Configurations* recommended by the JCT-VC [13] for the *All-Intra* (AI), *Low-Delay* (LP) and *Random-Access* (RA) mode configurations. That recommendation specifies the use of four *Quantization Parameter* (QPs) (22, 27, 32 and 37) and a set of 18 test sequences classified in five classes, A to E, which cover a wide range of resolutions from the largest (2560×1600 pixels) to the smallest (416×240 pixels), and frame rates from 60 fps to 24 fps. All the sequences use 4:2:0 chroma subsampling and a bit-depth of 8 bits:

- Class A (2560×1600 pixels): *Traffic* and *PeopleOnStreet*.
- Class B (1920×1800 pixels): *Kimono, ParkScene, Cactus, BQTerrace* and *BasketballDrive*.

 ISBN: 978-84-608-6082-2

- Class C (832×480 pixels): *RaceHorsesC, BQMall, PartyScene* and *BasketballDrill*.
- Class D (416×240 pixels): *RaceHorses, BQSquare, BlowingBubbles* and *BasketballPass*.
- Class E (1280×720 pixels): *FourPeople, Johnny* and *KristenAndSara*.

4.2. Metrics

The rate-distortion/complexity analysis was evaluated in terms of *Computational Complexity Reduction* (CCR) and R-D performance for all open-source HEVC codecs under study, and both of them were compared to the HM codec results used as reference. For the CCR measure, the *Time Saving* (TS) metric was computed following the Equation (1):

$$T.Saving(\%) = \frac{Enc.Time(HEVCcodec) - Enc.Time(HM)}{Enc.Time(HM)} \cdot 100 \quad (1)$$

Regarding the R-D performance, the average *Peak Signal to Noise Ratio* (PSNR) metric was calculated for each luma ($PSNR_Y$) and chroma components ($PSNR_U$, $PSNR_V$) and for the full number of frames of each sequence. In order to obtain a global quality measure, the average PSNR of the three components, denoted as $PSNR_{YUV}$, was also computed. As mentioned above, the chroma subsampling format of the test sequences was 4:2:0, so the well-known $PSNR_{YUV}$ according to Equation (2) was used, which applies weight pondering to the PSNR obtained for each component. Those weights are recommended by the JCT-VC in [14] and they are considered a fair representation of the visual quality, which is more sensitive to the luminance stimulus than the chrominance.

$$PSNR_{YUV}(dB) = \frac{6 \cdot PSNR_Y + PSNR_U + PSNR_V}{8} \quad (2)$$

The $PSNR_{YUV}$ for the four QPs were used for the computation of the R-D performance by using the *Bjøntegaard Delta Rate* (BDR) metric defined by ITU [15, 16] and recommended by the JCT-VC. The BDR provides the average difference between the R-D curves measured as a percentage of bit rate that is necessary to increase or decrease to achieve the same PSNR quality in both curves. In our simulation, a positive BDR means the encoded bit rate using the open-source HEVC codec under study is higher than the bit rate obtained with HM, thus that is denoted as the penalty in terms of bit rate.

 ISBN: 978-84-608-6082-2

4.3. Simulation Results

Table 2 shows the experimental results of open-source HEVC codecs under study compared with the HM 16.6 reference software for the AI configuration. It can be observed that HomerHEVC codec obtain the highest time savings (98.42%) while increasing the bit rate penalty by around 35%. On the contrary, x265 codec obtains the lowest bit rate penalty (11%) with a considerable average same saving (70%). The results obtained by Kvazaar report much better time savings of over 25% compared with x265, and a bit rate increase lower than 2.5% in terms of BDR, which is a very good balance between speed-up and rate penalty. Finally, the F265 codec shows a huge computational complexity reduction of over 94%, while increasing the bit rate penalty by around 69%.

Table 2: Time Savings (TS) and Coding Efficiency (BDR) results for AI configuration.

Classification	Sequence	Frames	**HomerHEVC**		**Kvazaar**		**x265**		**F265**	
			TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)
Class A (2560×1600)	Traffic	150	-98,36	34,9	-95,73	13,0	-71,03	7,4	-93,70	76,4
	PeopleOnStreet	150	-98,30	37,2	-95,79	12,7	-71,94	7,1	-93,81	77,9
Class B (1920×1080)	Kimono	240	-98,50	42,3	-95,97	13,4	-68,58	11,3	-93,24	76,1
	ParkScene	240	-98,44	25,5	-95,53	12,0	-69,22	7,4	-93,95	45,0
	Cactus	500	-98,46	35,2	-95,68	14,5	-69,83	10,9	-93,99	63,1
	BasketballDrive	500	-98,51	53,7	-95,98	15,6	-70,04	13,7	-93,64	104,8
	BQTerrace	600	-98,48	34,5	-95,84	13,5	-71,30	9,3	-94,32	64,3
Class C (832×480)	BasketballDrill	500	-98,42	40,6	-95,50	15,8	-71,65	10,4	-94,05	86,3
	BQMall	600	-98,37	36,0	-95,59	13,0	-71,14	10,3	-94,09	72,6
	PartyScene	500	-98,37	24,0	-95,12	11,3	-72,06	6,4	-95,04	39,1
	RaceHorses	300	-98,44	26,3	-95,34	11,3	-69,71	8,3	-94,14	46,3
Class D (416×240)	BasketballPass	500	-98,34	31,6	-95,25	13,4	-70,46	19,9	-94,27	62,5
	BQSquare	600	-98,44	19,9	-95,25	12,0	-71,97	11,4	-95,17	39,4
	BlowingBubbles	500	-98,36	23,7	-94,97	12,0	-70,95	10,9	-95,17	36,8
	RaceHorses	300	-98,37	25,5	-95,05	12,6	-68,86	15,1	-94,54	48,6
Class E (1280×720)	FourPeople	600	-98,39	42,1	-96,27	13,0	-71,40	10,0	-93,52	96,1
	Johnny	600	-98,57	49,5	-96,79	16,0	-70,61	15,3	-93,30	109,6
	KristenAndSara	600	-98,54	41,8	-96,81	15,0	-70,78	12,7	-93,31	104,0
Class A			-98,33	36,07	-95,76	12,84	-71,49	7,22	-93,76	77,14
Class B			-98,48	38,21	-95,80	13,81	-69,80	10,52	-93,83	70,66
Class C			-98,40	31,71	-95,39	12,84	-71,14	8,86	-94,33	61,11
Class D			-98,38	25,16	-95,13	12,49	-70,56	14,34	-94,79	46,82
Class E			-98,50	44,49	-96,62	14,64	-70,93	12,68	-93,37	103,25
AVERAGE			**-98,42**	**34,67**	**-95,69**	**13,33**	**-70,64**	**10,99**	**-94,07**	**69,39**

Table 3 shows the experimental results of open-source HEVC codecs under study compared with the HM 16.6 reference software for the LP configuration. It can be observed that all open-source HEVC codecs show a huge computational complexity reduction of over 95% (except F265 codec), while increasing the bit

rate penalty by over 60% which is not a very good balance between computational complexity reduction and rate penalty. High time savings are obtained by the open-source HEVC codecs using several suboptimal and reduced set of prediction modes in the picture partitioning.

Table 3: Time Savings (TS) and Coding Efficiency (BDR) results for LP configuration.

Classification	Sequence	Frames	**HomerHEVC**		**Kvazaar**		**x265**		**F265**	
			TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)
Class A (2560×1600)	Traffic	150	-98,56	70,1	-96,39	66,5	-95,70	47,9	-41,90	80,4
	PeopleOnStreet	150	-98,93	38,4	-96,28	39,5	-96,01	37,4	-60,30	53,2
Class B (1920×1080)	Kimono	240	-98,78	51,5	-96,39	49,0	-95,69	37,2	-50,91	64,7
	ParkScene	240	-98,66	52,3	-96,20	55,2	-95,49	55,5	-46,91	61,3
	Cactus	500	-98,77	74,0	-96,35	61,1	-95,93	62,4	-49,92	91,5
	BasketballDrive	500	-98,89	50,1	-96,41	53,6	-95,86	48,4	-54,34	82,7
	BQTerrace	600	-98,69	165,9	-96,21	99,4	-95,37	78,7	-46,23	100,8
Class C (832×480)	BasketballDrill	500	-98,87	70,8	-96,13	46,1	-96,13	56,5	-55,25	105,5
	BQMall	600	-98,77	61,3	-96,00	63,9	-96,05	58,6	-51,71	70,1
	PartyScene	500	-98,87	109,4	-95,91	53,7	-95,63	84,1	-57,24	71,7
	RaceHorses	300	-99,06	42,8	-96,31	50,4	-96,20	37,6	-63,39	52,0
Class D (416×240)	BasketballPass	500	-98,91	40,4	-96,14	40,6	-96,16	40,6	-60,87	55,3
	BQSquare	600	-98,71	237,2	-95,75	99,5	-95,19	127,7	-53,66	91,9
	BlowingBubbles	500	-98,77	86,6	-95,73	56,6	-95,66	81,0	-56,44	69,1
	RaceHorses	300	-99,00	43,3	-96,19	49,9	-96,01	46,1	-64,78	45,0
Class E (1280×720)	FourPeople	600	-98,49	103,1	-96,85	67,6	-96,56	56,9	-33,86	125,2
	Johnny	600	-98,43	167,7	-97,10	120,3	-96,37	76,8	-29,42	161,7
	KristenAndSara	600	-98,52	125,1	-97,15	88,8	-95,99	59,1	-33,87	135,7
Class A			-98,75	54,23	-96,33	52,98	-95,85	42,64	-51,10	66,75
Class B			-98,76	78,75	-96,31	63,64	-95,67	56,43	-49,66	80,21
Class C			-98,89	71,07	-96,09	53,52	-96,00	59,20	-56,90	74,82
Class D			-98,85	101,90	-95,96	61,67	-95,75	73,86	-58,94	65,32
Class E			-98,48	131,97	-97,03	92,23	-96,31	64,28	-32,38	140,84
AVERAGE			**-98,76**	**88,33**	**-96,30**	**64,53**	**-95,89**	**60,70**	**-50,61**	**84,31**

Table 4 shows the experimental results of open-source HEVC codecs under study compared with the HM 16.6 reference software for the RA configuration. It can be observed that Divx265 codec obtain the highest time savings (99.19%) while increasing the bit rate penalty by around 64%. On the contrary, x265 codec obtain the lowest bit rate penalty (55%) with similar time saving (94%) compared with Divx265. Kvazaar and HomerHEVC codecs shows an excessive bit rate penalty of over 100% with similar time saving results (around 96%) compared with Divx265 and x265. Finally, F265 presents the worst balance between computational complexity reduction and rate penalty. It is obvious that these open-source HEVC codecs use suboptimal decisions in order to speed up the encoding process at the expense of an increase in bit rate penalty.

 ISBN: 978-84-608-6082-2

Table 4: Time Savings (TS) and Coding Efficiency (BDR) results for RA configuration.

Classification	Sequence	HomerHEVC		Divx265		Kvazaar		x265		F265	
		TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)	TS (%)	BDR (%)
Class A (2560×1600)	Traffic	-98,60	130,6	-99,33	53,3	-96,95	99,0	-94,24	41,1	-58,11	201,9
	PeopleOnStreet	-98,89	60,3	-99,32	50,9	-96,61	56,0	-94,38	40,3	-69,56	139,1
Class B (1920×1080)	Kimono	-98,74	109,9	-99,12	41,9	-96,78	84,8	-93,49	48,1	-63,56	192,7
	ParkScene	-98,65	114,7	-99,11	53,6	-96,75	88,7	-93,79	47,7	-61,11	157,6
	Cactus	-98,76	130,4	-99,04	55,2	-96,77	103,4	-94,45	64,1	-61,74	221,0
	BasketballDrive	-98,86	99,6	-98,91	45,0	-96,81	87,0	-94,15	63,7	-63,69	199,8
	BQTerrace	-98,72	221,8	-98,99	73,5	-96,65	189,7	-93,69	69,8	-58,53	230,4
Class C (832×480)	BasketballDrill	-98,84	114,9	-99,29	63,5	-96,64	70,5	-94,91	47,7	-65,74	216,2
	BQMall	-98,74	132,1	-99,22	76,0	-96,51	117,0	-94,42	67,1	-60,89	205,8
	PartyScene	-98,80	178,0	-99,19	71,1	-96,18	98,9	-94,21	60,7	-65,26	152,8
	RaceHorses	-99,00	68,9	-99,19	61,2	-96,51	76,0	-94,63	46,2	-71,74	158,6
Class D (416×240)	BasketballPass	-98,87	72,0	-99,21	59,6	-96,56	64,1	-94,60	47,1	-68,75	143,5
	BQSquare	-98,63	320,7	-99,24	106,0	-96,07	196,8	-93,52	83,1	-61,39	187,3
	BlowingBubbles	-98,70	155,8	-99,17	77,8	-96,14	101,9	-94,29	64,7	-64,21	157,7
	RaceHorses	-98,93	69,9	-99,19	69,5	-96,41	75,1	-94,34	49,1	-72,31	150,3
Class E (1280×720)	FourPeople	-98,59	162,3	-99,32	51,3	-97,53	89,7	-95,28	45,5	-53,17	276,7
	Johnny	-98,58	257,1	-99,23	72,8	-97,71	141,6	-94,90	60,2	-51,10	341,5
	KristenAndSara	-98,62	190,3	-99,25	65,7	-97,76	116,6	-94,74	56,3	-52,49	307,8
Class A		-98,75	95,45	-99,32	52,10	-96,78	77,49	-94,31	40,70	-63,84	170,50
Class B		-98,75	135,28	-99,04	53,83	-96,75	110,69	-93,91	58,68	-61,73	200,29
Class C		-98,85	123,50	-99,22	67,96	-96,46	90,60	-94,54	55,42	-65,91	183,37
Class D		-98,78	154,58	-99,20	78,22	-96,30	109,47	-94,19	60,99	-66,67	159,71
Class E		-98,60	203,24	-99,27	63,25	-97,67	115,98	-94,97	54,00	-52,25	308,66
AVERAGE		**-98,75**	**143,85**	**-99,19**	**63,77**	**-96,74**	**103,15**	**-94,33**	**55,69**	**-62,41**	**202,26**

5. Conclusions

This paper presents a rate-distortion/complexity analysis of the several open-source HEVC codecs using objective measures of assessment in order to analyse their real capabilities. For AI configuration Kvazaar and x265, obtain a very good balance between speed-up and rate penalty. However, for LP and RA configurations, it can be observed that most open-source HEVC codecs under study shows a huge computational complexity reduction of over 95%, while increasing the bit rate penalty in an excessive way (by over 100%) which is not a very good balance between computational complexity reduction and rate penalty. Finally, x265 could be selected due to its cutting-edge status among the open-source HEVC encoders under study.

Acknowledgements

This work was jointly supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER funds) under the project TIN2015-66972-C5-2-R, and by the Spanish Ministry of Education, Culture and Sports under the grant FPU13/04601.

References

[1] ISO/IEC AND ITU-T, *Advanced Video Coding for Generic Audiovisual Services, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (AVC)*, February 2012.

[2] ISO/IEC AND ITU-T, *High Efficiency Video Coding (HEVC). ITU-T Recommendation H.265 and ISO/IEC 23008-2 (version 3)*, April 2015.

[3] J.-R. OHM, G. J. SULLIVAN, H. SCHWARZ, THIOW KENG TAN, AND T. WIEGAND, *Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC)*, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1669–1684, December 2012.

[4] *Joint Collaborative Team on Video Coding Reference Software, ver. HM 16.6*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[5] *x265*. [Online]. Available: http://x265.org/

[6] *F265*. [Online]. Available: http://vantrix.com/f-265/

[7] Kvazaar HEVC encoder. [Online]. Available: https://github.com/ultravideo/kvazaar

[8] *Ultra video group* [Online]. Available: http://ultravideo.cs.tut.fi/

[9] *HomerHEVC Encoder*. [Online]. Available: https://github.com/jcasal-homer/HomerHEVC

[10] *DivxHEVC encoder*. [Online]. Available: http://labs.divx.com/divx265

[11] G. J. SULLIVAN, J. R. OHM, W. J. HAN, AND T. WIEGAND, *Overview of the High Efficiency Video Coding (HEVC) Standard*, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649-1668, December 2012.

[12] IL-KOO KIM, JUNGHYE MIN, TAMMY LEE, WOO-JIN HAN, AND JEONGHOON PARK, *Block Partitioning Structure in the HEVC Standard*, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1697-1706, December 2012.

[13] F. BOSSEN, *Common Test Conditions and Software Reference Configurations*, document JCTVC-L1100, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), 12th Meeting: Genève, CH 14 – 23, January 2013.

[14] G. J. SULLIVAN, KOOHYAR MINOO, *Objective Quality Metric and Alternative Methods for Measuring Coding Efficiency*, document JCTVC-H0012, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), 8th Meeting: San Jose, CA, USA, 1 – 10, February 2012.

[15] G. BJØNTEGAARD, *Calculation of Average PSNR Differences between RD-Curves*, ITU-T SG16 Q.6 Document VCEG-M33, Austin, US, April 2001.

[16] G. BJØNTEGAARD, *Improvements of the BD-PSNR Model*, ITU-T SG16 Q6 Document VCEG-AI11, 35th VCEG Meeting, July 2008.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Chirality at the Nanoscale

Ignacio L. Garzón
Instituto de Fisica
Universidad Nacional Autónoma de México
51 Ciudad de México, México

email: garzon@fisica.unam.mx

Abstract

In this work, a briefly review of the most recent results on the geometric quantification of chirality through a calculation of the Hausdorff chirality measure for several well-known bare and ligand-protected gold clusters will be presented. In addition, results, based on density functional theory calculations, on the enantiospecific adsorption of the cysteine amino acid on chiral gold clusters will also be discussed.

Key words: chirality, clusters, enantiospecific adsorption

1. Introduction

Chirality has been found as a relevant property of nanomaterials, including ligand-protected metal clusters and nanorods. This property is not only crucial in nanotechnology developments related with asymmetric catalysis and chiroptical phenomena, but also generates fundamental questions on the existence of chirality at the nanoscale. In fact, x-ray total structure determination, electron diffraction studies, NMR and circular dichroism spectroscopies, as well as theoretical calculations performed on gold clusters protected with thiolate or phosphine ligands have confirmed the existence of chiral structures in the size range of 18-144 Au atoms. In this work [1], we realize a comparative analysis of the degree or amount of chirality existing in chiral ligand-protected gold clusters (LPGC), through a geometric quantification, using the Hausdorff chirality measure (HCM).

The interaction of biological molecules like chiral amino acids with chiral metal clusters is becoming an interesting and active field of research because of its potential impact in, for example, chiral molecular recognition phenomena. In

 ISBN: 978-84-608-6082-2

particular, the enantiospecific adsorption (EA) of cysteine (Cys) on a chiral Au_{55} cluster was theoretically predicted a few years ago [2,3]. In this work, we also present theoretical results, based on density functional theory, of the EA of non-zwitterionic cysteine interacting with the C_3-Au_{34} chiral cluster [4], which has been experimentally detected in gas phase, using trapped ion electron diffraction.

2. Results

The geometric quantification of chirality in LPGCs across the size range of 18-144 Au atoms was performed through the HCM approach, using the cluster structural information obtained from XRC and TEM measurements, as well as from DFT calculations. The calculated HCM values assign a quantitative measurement to the chirality existing in LPGC, including those corresponding to the cluster core, its protecting ligand arrangement, and their combination. These HCM values allow for the comparison between the chirality existing in the cluster core and in the ligand arrangement in order to gain insight into the origin of chirality. In addition, it is also possible to compare the index of chirality of LPGC among different sizes. The results presented in this work [1] indicate that the HCM values are consistent with those known trends on the origin of chirality obtained from symmetry considerations and CD measurements discussed previously for some chiral clusters. For example, the chirality of the $Au_{38}(PET)_{24}$ and Au_{102}(p-MBA)$_{44}$ clusters has been mainly attributed to the chiral arrangement of the ligands protecting a nearly symmetric (achiral) core, whereas in the $Au_{20}(PPh_3)_4$ cluster, the chirality has been associated with the existence of an intrinsically chiral Au_{20} core, composed of an icosahedral Au_{13} core surrounded by a helical Y-shaped Au_7 motif. The HCM values obtained for these chiral clusters not only confirm these qualitative trends, but also provide a framework for a systematic quantification of chirality. Moreover, our results have also shown the capability of the HCM approach to provide new information on the chirality of clusters that was not mentioned in their initial reports describing their original synthesis and structural characterization. In this respect, the HCM approach was also useful to provide insights into the chiroptical properties of LPGC. In fact, we consider that this capability would be an adequate alternative for an initial understanding of chirality in the larger LPGCs to be synthesized, where first principles calculations of the CD spectrum will be scarce due to the high computational cost involved. In summary, our results show the clear advantages for a systematic quantification of chirality in LPGCs through the HCM approach, both in terms of its consistency with known trends and the capability to provide new insights. In particular, it has been shown that the Au-S interface plays a preponderant role in the overall chirality of most of the LPGC under study. Future extensions of this work, that would provide further validation of the geometrical quantification of chirality, are the search for systematic correlations between the HCM values and physical or chemical observables like

 ISBN: 978-84-608-6082-2

optical activity, and the strength of the enantiomer-specific interaction between LPGC and other chiral systems [1].

Our results also show that the adsorption energy of the amino acid depends on which enantiomers participate in the formation Cys-Au_{34} chiral complex [4]. EA was obtained in the adsorption modes where both the thiol, and the thiol-amino functional groups of Cys are adsorbed on low-coordinated sites of the metal cluster surface. Similarly to what was obtained for the Cys-Au_{55} chiral complex, in the present work, it is found that the EA is originated from the different strength and location of the bond between the COOH functional group and surface Au atoms of the Au_{34} chiral cluster. Calculations of the vibrational spectrum for the different Cys-Au_{34} diastereomeric complexes predict the existence of a *vibro-enantiospecific effect*, indicating that the vibrational frequencies of the adsorbed amino acid depend on its handedness.
The present results provide a useful example of how the structural, vibrational, and energetic properties of a chiral complex (chiral molecule interacting with a chiral cluster) depend on the enantiomers of its constituents. Although this phenomenon was previously reported for the structural and energetic properties of the Cys-Au_{55} complex [2.3], in this work, it is also predicted for the vibrational frequencies of the Cys-Au_{34} chiral complex. In order to experimentally confirm the enantiospecific phenomena mentioned above, the development of novel and accurate enantioseparation methods is necessary, as well as precise a control of the synthesis and composition of the chiral complexes.

3. References

[1] J. J. PELAYO, R. L. WHETTEN, I. L. GARZÓN, *Geometric quantification of chirality in ligand-protected metal clusters*, J. Phys. Chem. C **119** (2015) 28666.

[2] X. LÓPEZ-LOZANO, L. A. PÉREZ, I. L. GARZÓN, *nantiospecific adsorption of chiral molecules on chiral gold clusters,* Phys. Rev. Lett. **9** (2006) 233401.

[3] L. A. PÉREZ, X. LÓPEZ-LOZANO, I. L. GARZÓN, *Density functional study of the cysteine adsorption on u nanoclusters,* Eur. Phys. J. D **52** (2009) 123.

[4] J. J. PELAYO, I. VALENCIA, G. DÍAZ, X. LÓPEZ-LOZANO, I. L. GARZÓN, *Enantiospecific adsorption of cysteine on a chiral Au34 cluster,* Eur. Phys.
J. D **9** (2015) 277.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Hedonic and spatial analyses applied to the massive assessment of real estate appraisals performed by appraisal companies in the Province of Valencia (Spain)

Natividad Guadalajara[1] and Miguel Ángel López[2]

[1] *Centro de Ingeniería Económica, Universitat Politècnica de València, Spain*

[2] *Universitat Politècnica de València, Spain*

emails: nguadala@omp.upv.es, mangel.lopez@bde.es

Abstract

Traditionally, hedonic models have been widely used in mass assessments of properties, but may present spatial autocorrelation problems. By using the property appraisal registers of four Spanish appraisal companies, several spatial and hedonic models were developed. Better results were obtained by the spatial models.

Key words: ordinary least squares, spatial econometric model, housing values, appraisal companies, mass appraisal

1. Introduction

Real estate valuations are a complex matter because of the high heterogeneity of the characteristics of immovable properties, which makes performing a massive valuation system difficult. Furthermore, real estate appraiser's independence has been periodically scrutinized in not only Spain, but also in many other countries, whenever a plunge in real estate prices occurs.

In Spain, the significance of the mortgage lending market, and the real estate industry in general, is very marked. In 2013 the real estate sector peaked at 13% of Spain's GDP. According to the "Boletin Económico" of the Banco de España (Bank of Spain) [1], the total amount of mortgage lending granted for home

 ISBN: 978-84-608-6082-2

acquisition and rehabilitation continued to rise until 2010, when it peaked at the largest amount of 663 billion €. In 2013 total mortgage loans reached 65% of the GDP, which is much higher than in Italy (22%), France (41%) or Germany (46%), but below the Netherlands (107%), Denmark (101%) or the United Kingdom (85%). It is also important to consider the stock of more than 6 million mortgages by the end of 2012.

In order to determine the proper value of real estate collaterals, appraisals should be considered an important issue in property assessments. Appraisal values are critical in banking systems and in the mortgage lending market.

When a mortgage originates, an appraisal is necessary to determine a very relevant ratio for risk control and management purposes: the loan-to-value ratio (LTV Ratio) is a lending risk assessment ratio that financial institutions and other lenders examine before approving mortgages. Typically, assessments with high LTV ratios are generally seen as being at higher risk. It represents the ratio of the first mortgage lien as a percentage of the total appraised value of real property.

Although it is not difficult to calculate the amount to be loaned, the property value is not estimated directly and should be appraised by a professional real estate valuer.

Likewise, prudential regulation requires the periodically assessment of the mortgages portfolio in order to determine the capital requirements level: this depends on the type of collaterals and the LTV ratio level: e.g. if loans are collateralized by residential property, and when the LTV is lower than 80%, capital requirements are lower than if LTV >80%.

Moreover, regarding non performing loans portfolios or foreclosed assets, it is necessary to appraise property to evaluate the expected losses of loans or impairment of property.

Finally, the properties and premises that correspond to banks are to be accounted at a fair value, and then an appraisal is required.

Having accurate appraisals provides valuable information when mortgage loan procedures arise, and in cases of impairment, losses are minimized (the appraisal should enclose information about the probability of default). Depending on the information cost and the losses reduction that an appraisal facilitates, appraisal market prices should be optimized. This is the main reason to develop massive valuation models that will lead to efficient fast immovable property valuations.

 ISBN: 978-84-608-6082-2

Traditional linear hedonic models were very popular at the beginning of the 1960s, but have received plenty of criticism because of specification inaccuracies, such as omitting linearity principles [2] and multicollinearity problems, non linearity and heteroscedasticity [3]. These mistakes means that not all the predictor variables have the appropriate sign and residual values in the scatter plot graph, show a trend along predicted values, and are not compatible with the principle of independent and identically distributed errors [4].

Spatial interaction produces proximity or neighborhood effects on different scales. This problem was recognized by Student in 1914 and was solved by spatial econometrics. Its development took place in the 1980s and 1990s thanks to the implementation of geographical information with high availability and data access, and also to the development of Geographic Information Systems (GIS) and specific software for spatial data analyses [5]. In GIS, data are geo-referenced: latitude and longitude or coordinates XUTM and YUTM.

Autocorrelation or spatial dependence implies that the value of a variable is conditioned by the value of that variable in one region or more adjacent ones. Therefore, the First Law of Geography states that: "everything is related to everything else, but near things are more related than distant things" [6].

Spatial regression models have been used in different works of massive real estate valuations: Herath and Maier (2013) used a sample of 1656 bid prices apartments in Vienna in 2009-2010 [7]; Zang, Du, Geng, Liu and Huang (2014) analyzed 236 sales of commercial properties in China [8]; Zoopi, Argiolas and Lai (2015), who analyzed those factors with an influence on housing values in Cagliari in 2012, used a sample of 304 apartments [9].

The aim of this paper was to model the appraised values of multifamily apartments in the province of Valencia (Spain). We used 5551 data for 2014 provided by four appraisal companies to develop hedonic and spatial regression models by maximum likelihood.

The reason for selecting the province of Valencia is because it has several features that make it ideal as a pattern of the main characteristics of the Spanish residential market: Valencia is a big city with 800 000 inhabitants. The province is located on the littoral with high tourist influx towns such as Oliva, Cullera and Gandia. Inland there are towns with industrial or agricultural activity, such as Requena, Onteniente and Alzira.

2. Data

The regression model and the successive tests analysis are based on a database of 5551 records of multifamily housing in the province of Valencia. The source of these data corresponds to the appraisals completed during 2014 by four Spanish Appraisal Companies following the information requirements from Order ECO / 805/2003, which include Spanish real estate valuation methodologies. The houses valuation breakdown for each company is: 1313 Company S0; 3348 Company S1; 362 Company S2 and 528 by Company S3.

Apart from the appraised value (in euros), other variables are selected, such as area (in square meters) and the age of the house in years (corrected with the time from the last complete overhaul, wherever appropriate).

Table 1 shows the descriptives of the selected quantitative variables.

Table 1. Characteristics of properties

	Value (euros)	Value per square meter (euros/m^2)	Area (m^2)	Age (Years)
Mean	113,669.9	1078.7	101.7	36.6
Standard desv	94,418.6	607.5	33.4	22.6
Minimum	11,865.0	155.0	30.0	1.0
Maximum	1,600,332.0	5673.5	639.0	195.0
Mode	91,403.0	1269.5	72.0	45.0
Median	90,750.0	926.4	98.0	40.0

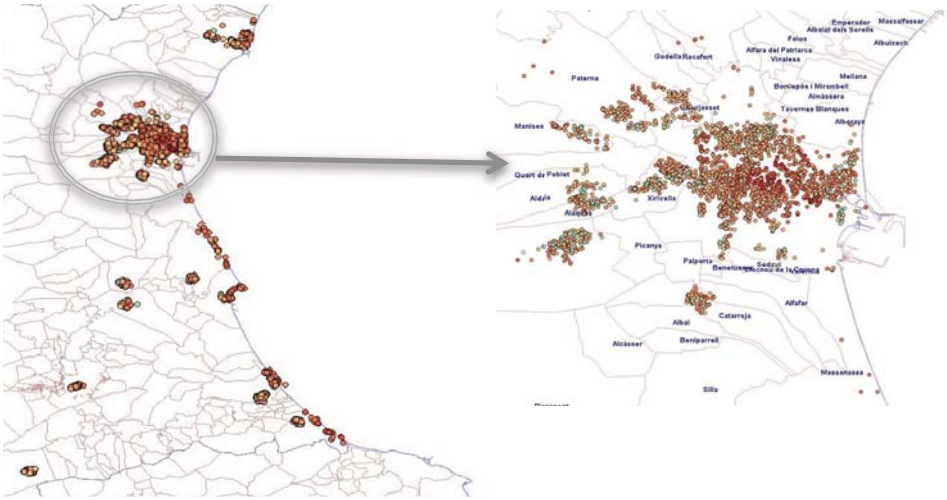
Other available house characteristics are: swimming pool, air conditioning, building quality, preservation state, number of parking lots, type of floor, m^2 terrace, etc.

The "jerarquía" variable was also selected, which is obtained from the Dirección General del Catastro, and refers to the value of land considered for cadastral purposes. "Jeraquía" ranges from 1 (top land value) to 67 (lower value).

Finally, the housing spatial location is defined by the geographic coordinates longitude and latitude. Graph 1 shows the value map of the houses near the city of Valencia city to be analyzed, and others located in northern areas (Sagunto) or in southern areas (Gandia, Onteniente, etc.). As we can see, there are areas with a higher value (dark red) in the city center of Valencia and the former Turia riverbank, with the highest values (dark red) similarly to some areas close to the

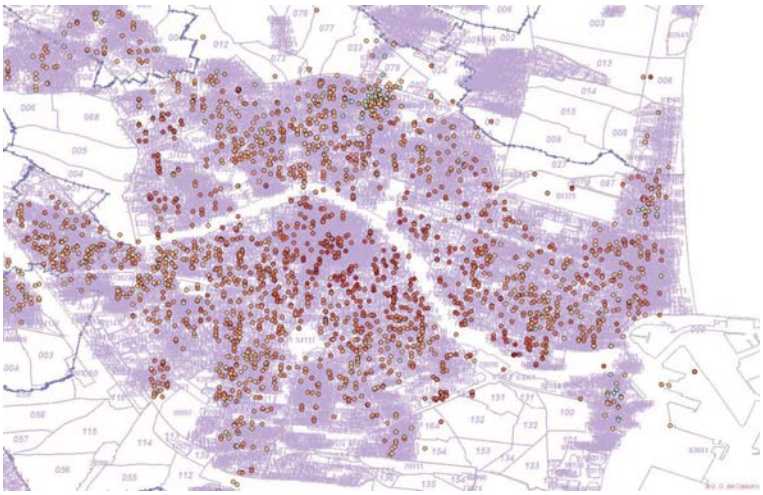
seaside. In the metropolitan area of Valencia, values are lower, but with some heterogeneity.

Graph 1. Map of the value of the properties located in the province of Valencia



More details in Graph 2: housing values in downtown Valencia.

Graph 2. Housing Map Values in downtown Valencia



3. Methodology

Following other works [7,9,10], we firstly estimated the Ordinary Least Squares (OLS) Hedonic Regression Model, and then the spatial econometric models.

The Hedonic Regression Model expression is:

 ISBN: 978-84-608-6082-2

$$Y_i = \alpha + \beta_j * X_{ji} + \varepsilon$$

where: Yi = House value i
Xji = Vector of the j explanatory variables for property i
ε = Error term

This regression model was based in the following hypothesis: E[ε]=0; homoscedasticity; normality of residuals; linearity and absence of multicollinearity.

We considered the predicted variables to be: total house value and per square meter house value, both logarithmically and as referred to in several works [4,7,11].

The predictor variables we considered were as follows:

a) Houses structural characteristics:
   Quantitative:
   Ln [size (m^2)]^2: the squared Ln of the housing area.
   Ln [age (years)]^2: the squared Ln of the housing age (corrected with the time from the last complete overhaul, wherever appropriate).
   Qualitative:
   Swimming pool: if there is a swimming pool in the condominium (1= yes; 0=no)
   Air conditioning: Air conditioning in the apartment (1= yes; 0=no)
   Construction quality: Building quality (1= High quality; 0=rest)
   Preservation: Housing state of preservation (1= good or very good; 0=rest)
   Lift: The building has an elevator (1= yes; 0=no)
b) Socio-economic characteristics:
   L *Jerarquía*: the ln of the *jerarquía* value. It depends on the characteristics of the property's location.
c) Appraisal company: S0, S1, S2 and S3. They are four dummy variables. Their value is 1 if the appraisal was performed by company 0, and 0 otherwise.

The OLS model tests considered to analyze the significance of the regression coefficients were the Snedecor's F-test and the student's t-test. The R^2 and the log likelihood ratios were considered to compare the goodness of fit of the models, as well as the Akaike (AIC) and Schwartz criteria (SC).

Besides the following were analyzed: the condition number of multicollinearity, the Jaque-Bera normality test, and the Breusch-Pagan and Koencker-Bassett heteroscedasticity tests.

 ISBN: 978-84-608-6082-2

After selecting the best hedonic model, we analyzed the spatial autocorrelation statistical parameters: Moran's I and Lagrange Multiplier (LM), developed according to the maximum likelihood theory [7,12]. Moran`s Index indicates if any type of spatial correlation appears, and the LM tests specify the possibility of analyzing the hypothesis of non spatial dependence caused by an omitted spatial lag or of spatially correlated errors. So the LM test checked two types of spatial correlation: spatial lag dependence ρ and spatial error, λ:

The Spatial Autoregressive Lag Model (SAR) is:

$$Y = \alpha + \beta_i * X_i + \rho * W_y + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

The Spatial Autoregressive Error Model (SEM) is.

$$Y = \alpha + \beta_i * X_i + \varepsilon$$
$$\varepsilon = \lambda * W_\varepsilon + \mu \qquad \mu \sim N(0, \sigma^2)$$

Where
ρ = is the spatial correlation coefficient
λ = is the error correlation coefficient
W = is the spatial weights matrix

This weights matrix is symmetric, transposed, positive, non stochastic and squared N*N, where N is the number of observations or immovable properties with the elements or weights (w_{ij}) that outline the interdependence intensity between each couple of housing values I and j. Diagonal values are always zero (w_{ii}=0), and this matrix can be estimated considering the distances between properties. The most popular to consider is the Euclidean metric [3], and this was applied in this work. However, there are many other possibilities, and sometimes weights are corrected with other variables, such as the importance index from each business area [8]**.** The software tool used was Geoda.

4. Results

The results of the two OLS linear regression selected models are shown in the second column in Tables 2 and 3 after considering the total housing values and the per square meter value, both logarithmically respectively.

Table 2. Comparison of the estimated results of the OLS regression and spatial econometric models for total values

 ISBN: 978-84-608-6082-2

	OLS	Spatial Lag (SAR)	Spatial Error (SEM)
Variable	Coefficient	Coefficient	Coefficient
House Characteristics			
Ln [size (m^2)]^2	0.09675***	0.09662***	0.09685***
Ln [age (years)]^2	-0.02432***	-0.02408***	-0.02381***
Swimming pool	0.17539***	0.17191***	0.17515***
Air conditioning	0.06854***	0.06896***	0.07111***
Construction Quality	-0.05679***	-0.05705***	-0.05548***
Preservation	0.12996***	0.12727***	0.13171***
Light	0.32552***	0.32361***	0.32532***
Socio-economic Characteristics			
Ln Jerarquia	-0.44025***	-0.43832***	-0.4399***
Valuation Companies			
S1	0.18702***	0.18240***	0.18579***
S2	0.47124***	0.45519***	0.47242***
S3	0.35080***	0.32931***	0.35216***
Constant	10.60451***	10.19084***	10.59464***
N	5,551	5,551	5,551
R^2	0.8076	0.8088	0.81045
F-statistic	2118.86		
ρ		0.03627***	
λ			0.09166***
Log likelihood	-888.248	-877.134	-864.097
Akaike info criterion (AIC)	1800.5	1780.27	1752.19
Schwarz (SC)	1879.96	1866.35	1831.65
Likelihood Ratio Test (LR test)		22.2270***	48.3017***

*** p-value <0.001

Table 3. Comparison of the estimated results of the OLS regression and the spatial econometric models for the per square meter values

	OLS	Spatial Lag (SAR)	Spatial Error (SEM)
Variable	Coefficient	Coefficient	Coefficient
House Characteristics			
Ln [size (m^2)]^2	-0.01204***	-0.01148***	-0.01184***
Ln [age (years)]^2	-0.0245***	-0.02385***	-0.02392***
Swimming pool	0.1727***	0.16697***	0.17265***
Air conditioning	0.0681***	0.07147***	0.07085***

 ISBN: 978-84-608-6082-2

Construction Quality	-0.0521***	-0.05327***	-0.05107***
Preservation	0.1301***	0.12650***	0.13202***
Light	0.3237***	0.32339***	0.32316***
Socio-economic Characteristics			
Ln Jerarquia	-0.44856***	-0.44466***	-0.44801***
Valuation Companies			
S1	0.1864***	0.18113***	0.18503***
S2	0.4835***	0.45311***	0.48515***
S3	0.3490***	0.32161***	0.35038***
Constant	8.3459***	7.9407***	8.33294***
N	5,551	5,551	5,551
R^2	0.7141	0.71646	0.71796
F-statistic	1257.92***		
ρ		0.05597***	
λ			0.09360***
Log likelihood	-922.017	-903.701	-896.817
Akaike info criterion (AIC)	1868.03	1883.4	1817.63
Schwarz (SC)	1947.49	1919.48	1897.09
Likelihood Ratio Test (LR test)		22.2270***	50.4006***

*** p-value <0.001
*** p-value <0.001

Both models showed a good fit level $R^2 = 0.81$ when considering the total properties value, and $R^2 = 0.71$ when considering the per square meter value. All the variables coefficients were significant at the 99% level of confidence.

The multicollinearity, heteroscedasticity, non normality and spatial dependence tests for each model are shown in Table 4.

Table 4. Test of multicollinearity, heteroscedasticity, non-normality and spatial dependence en OLS

	Ln Total_Value		Ln per_Sq_mtr Value	
Tests	DF	Value	DF	Value
Multicollinearity condition number		38.599		38.599
Heteroscedasticity				
Breusch-Pagan	11	406.63***	11	406.09***
Koencker-Bassett	11	278.06***	11	273.75***

 ISBN: 978-84-608-6082-2

Non-normality				
Jaque-Bera	2	199.097***	2	216.86***
Spatial dependence				
Moran's I		7.10***		7.25***
Lagrange Multiplier (lag)	1	22.326***	1	36.793***
Robust LM (lag)	1	0.678	1	2.261
Lagrange Multiplier (error)	1	49.48***	1	51.642***
Robust LM (error)	1	27.83***	1	17.11***

As we can see in Table 4, both models showed no multicollinearity problems, but heteroscedasticity problems were detected when the Breusch-Pagan and Koenker-Bassett tests were executed. The same problems appeared for normality errors according to the Jarque-Bera test.

The spatial dependence tests (Moran´s I, LMlag, RLMlag, LMerror, RLMerror) also showed spatial dependence problems, and better results were expected with the Spatial Error Model (SEM).

The third and fourth columns in Tables 2 and 3 show the Spatial Lag Models (SAR) and SEM compared with the OLS model.

5. Conclusions

The outputs of the OLS and ML models are strikingly similar, as other studies have indicated [9]**.**

Both the models obtained with the total value as the per square meter value gave significant differences among the four valuers, as evidenced by the significant coefficients of the variables associated with the companies. This indicates that, regardless of the housing and hierarchy characteristics, the company S0 valuations performed valuations downwardly, while the company S2 estimated values upwardly after comparing all the analyzed companies.

However, spatial dependence problems were detected in the OLS, which meant that spatial regression models had to be developed, as in other works on real estate valuations, and better results were obtained, as deduced from the R^2, Log likelihood, Akaike and Schwarz tests.

As in other works [7,10], but unlike others [9], SAR outperformed SLM, as evidenced by the Likelihood Ratio test.

 ISBN: 978-84-608-6082-2

6. References

[1] Boletín Económico: Banco de España - Indicadores económicos. Available at: http://www.bde.es/webbde/es/estadis/infoest/indeco.html [Accessed May 6, 2016].

[2] W.J. MCCLUSKEY, M. MCCORD, P.T. DAVIS, M. HARAN AND D. MCILHATTON, *Prediction accuracy in mass appraisal: a comparison of modern approaches*, Journal of Property Research. **30** (2013) 239-265.

[3] J. GUAN, D. SHI, J.M. ZURADA AND A.S. LEVITAN, *Analizing massive data sets: an adaptive fuzzy neural approach for prediction, with a real estate illustration*, Journal of Organizational Computing and Electronic Commerce. **24** (2014) 94-112.

[4] R.K. PACE AND O.W. GILLEY, *Using the Spatial Configuration of the Data to Improve Estimation*, Journal of Real Estate Finance and Economics. **14** (1997) 333-340.

[5] E. BELSKY, A. CAN AND I. MEGBOLUGBE, *A Primer on Geographic Information Systems in Mortgage Finance*, Journal of Housing Research. **9** (1998) 5-31.

[6] W. TOBLER, *A computer movie simulating urban growth in the Detroit region*, Economic Geography. **46** (1970) 234-240.

[7] S. HERATH AND G. MAIER, *Local particularities or distance gradient. What matters most in the case of the Viennese apartment market?*, Journal of European Real Estate Research. **6** (2013) 163-185.

[8] R. ZHANG, Q. DU, J. GENG, B. LIU AND Y. HUANG, *An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study*, Habitat International. **46** (2015) 196-205.

[9] C. ZOPPI, M. ARGIOLAS AND S. LAI, *Factors influencing the value of houses: Estimates for the city of Cagliary, Italy*, Land Use Policy. **42** (2015) 367-380.

[10] B. KIM AND T. KIM, *A Study on Estimation of Land Value Using Spatial Statistic: Focusing on Real Transaction Land Prices in Korea*, Sustainability. **8** (2016) 203-217.

[11] C. CHASCO AND J. LE GALLO, *The Impact of Objective and Subjective Measures of Air Quality and Noise on House Prices: A Multilevel Approach for Downtown Madrid*, Economic Geography. **89** (2012) 127-148.

[12] L. ANSELIN AND S. J. REY, *Modern Spatial Econometrics in Practice*, GeoDa Press, Chicago, 2014.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Optical properties of graphene quantum dots: from clusters to microstructures

P. Hawrylak*
Department of Physics, University of Ottawa,
Ottawa, K1N 6N5, Canada

emails: pawel.hawrylak@uottawa.ca

Abstract

We discuss here a theory of optical properties of graphene quantum dots with sizes from clusters to microstructures.

Key words: graphene quantum dots, electron-electron interactions, optical propoerties

1. Extended abstract

We present here a theory of optical properties of graphene quantum dots with sizes from clusters to microstructures. The electronic, optical and magnetic properties of graphene can be modify by engineering lateral size, shape, edge, number of layers and sublattice symmetry [1-4] of graphene quantum dots(GQD). Graphene quantum dots confine Dirac Fermions, both electrons and holes, with strength of their interaction controlled by both background dielectric constant and lateral confinement. Here we present new results describing the role of interactions in multi-exciton complexes composed of Dirac fermions in conduction and valence band, including theory of bi-excitons [4] in transient absorption spectra[5]. Building on our previous work [1-3] we describe the single-particle energy spectra using the tight-binding model based on P_z carbon orbitals. All direct and exchange two-body Coulomb matrix elements are computed using Slater P_z orbitals for on-site and nearest and next nearest neighbors and approximated for farther neighbors. The Coulomb interactions are screened by a dielectric constant of external medium and sigma electrons changing the ratio of Coulomb interactions to the tunneling matrix element. For a given GQD with a defined shape, size, edge, and dielectric constant we compute the fully self-consistent Hamiltonian of Hartree-Fock quasiparticles. The many body ground

 ISBN: 978-84-608-6082-2

and excited states are expanded in a finite number of electron-hole pair excitations from the Hartree-Fock ground state and computed using exact diagonalization technique with Hilbert spaces of the order of 10^6 configurations. This allows us to establish ground state and exciton and biexciton spectra and their Auger coupling. For colloidal GQDs the degeneracy of the top of the valence and bottom of conduction band leads to characteristic X and XX bands and two degenerate bright exciton states. Excitations from dark excitons to both excited exciton states and bi-excitons determine transient absorption spectrum and help us to identify possible XX-X-GS cascades. Comparison with experimental transient absorption spectra [5] will be made.

2. References

[1] O.VOZNYY, A.D.GUCLU, P.POTASZ, P.HAWRYLAK, Phys.Rev.**B83**, 165417 (2011).

[2] A.D. GUCLU, P. POTASZ, M. KORKUSINSKI AND P. HAWRYLAK,"*Graphene Quantum Dots*", Springer-Verlag (2014).

[3] P.HAWRYLAK, F PEETERS, K. ENSSLIN, Editors, *Carbononics–integrating electronics, photonics and spintronics with graphene quantum dots*, Focus issue, Physica status solidi (RRL)-Rapid Research Letters **10** (1), 11(2016).

[4] I.OZFIDAN, M. KORKUSINSKI AND P. HAWRYLAK, Phys.Rev.**B91**, 115314(2015).

[5] CENG SUN, FLORIAN FIGGE, I.OZFIDAN, M. KORKUSINSKI, XIN YAN, LIANG-SHI LI, PAWEL HAWRYLAK AND JOHN A. McGUIRE, NanoLetters **15**,5742(2015).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Study of influence of surface mass coefficient of free chloride in reinforced concrete using Spice code

Hidalgo, P., Sánchez-Pérez, J. F. and Alhama, I.

Network Simulation Research Group, Universidad Politécnica de Cartagena

emails: juanf.sanchez@upct.es, mdpht0@alu.upct.es, ivan.alhama@upct.es

Abstract

One of the main causes that triggers the steel corrosion of the reinforced concrete structures is the presence of chloride ions, which penetrates through diffusion process from the outside coming from marine environments. The threshold to switch on the corrosion process defines as the corresponding time from the moment when the aggressive agent begins to penetrate through the coating until it reaches the reinforced steel and depassivates. In this work, it is studied the influence of surface mass coefficient of free chlorides in this stage through the evolution of free chloride concentration.

Key words: Reinforced concrete; corrosion; modelling studies; free chloride concentration; partial differential equations

1. Introduction

Reinforced concrete arises when joining steel with the concrete, which produces good mechanical properties like great tensile or compressive strength. Concrete is a heterogeneous composite formed by the mixture of cement with water, resulting a material with porous structure. The steel used as reinforcement in the concrete serves to absorb tensile stress of the structure. The corrosion phenomena are associated with construction defects, changes in the service conditions of structures, or to the action of external aggressive agents. In these situations the mechanical strength of the structure is affected. The actuation of external aggressive agents depends on the environment where they are and on the penetration rate. These aggressive agents can be in gaseous or liquid state, or forming part of adjacent concrete floors. One of the main causes that can trigger the corrosion of reinforcing steel is the presence of chloride ions, either in raw materials of concrete or due to their penetration from the outside in marine environments. These ions produce specific breaks in the passive layer and therefore lead to localized corrosion.

 ISBN: 978-84-608-6082-2

Useful life means the time period in which the structure retains its geometric features, functionality and safety without unexpected costs repair or maintenance. The last part of the useful life is known as residual useful life, that is, while the corrosion process is developed. The initiation time of this last part is defined as the corresponding time from the moment when the aggressive agent begins to penetrate through the coating until it reaches the reinforced steel and depassivates.

In this paper, we will study this initiation time which depends on the property of concrete that determines the penetration rate of aggressive agents (chlorides), the porosity. The parameters that influence this property are the type and pore size, the moisture content and the composition of the aqueous phase [1].

2. Mathematical model

Based on conservation of mass and heat transfer, the non-linear mathematical model is formed by three main differential equations, related to the three species involved, chlorides (C_{cf}), humidity (h) and temperature (T). These are:

$$\frac{\partial C_{fc}}{\partial t} = \frac{\partial\left(w_e D_c^* \frac{\partial C_{fc}}{\partial x}\right)}{\partial x} + \frac{\partial\left(C_{fc} D_h^* \frac{\partial h}{\partial x}\right)}{\partial x} \quad (1)$$

$$\frac{\partial w_e}{\partial h}\frac{\partial h}{\partial t} = D_h^* \frac{\partial^2 h}{\partial x^2} \quad (2)$$

$$\rho_c c_q \frac{\partial T}{\partial t} = \lambda \frac{\partial^2 T}{\partial x^2} \quad (3)$$

where ρ_c is the concrete density, c_q, the specific heat and λ, the thermal conductivity.

These transient equations describe chloride penetration into concrete and moisture and heat diffusion through it, which depends on the boundary conditions as well as on the physical parameters of the concrete and the exposure time.

Boundary conditions are enforced to simulate seasonal variations in exposure conditions at the outer surface of the concrete. These are applied by assuming fluxes (convection) which depend on the difference between the environmental and the concrete surface values, this is by a second class or Neuman condition, by using suitable coefficients [2-3].

3. The network model

This is designed from the finite difference differential form of the governing equations 1 to 3 refer to each main species, chlorides, humidity and temperature, plus other seven that define the coefficients involved in those equations in term of the dependent variables. Boundary equations must be added to form the complete

 ISBN: 978-84-608-6082-2

model. The detailed rules for designing the model can be found in González-Fernández and Alhama [4]. The model, which can be simulated by means of circuit analysis software that uses Spice code, such as PSpice or NgSpice, [5-6], to provide the non-steady state solution, is based on the formal equivalence between finite-difference differential equations of the model and those of the physical process. To this end, the PDEs that form the governing equations are spatially discretized while time remains as a continuous variable in the model.

4. Simulation and results

To study the initiation time of residual useful life, we use the typical values of initial diffusivity of free chlorides, $4.5 \cdot 10^{-12}$ m^2/s, and initial diffusivity of humidity, $1.157 \cdot 10^{-10}$ m^2/s. Boundary conditions are defined by the surface mass coefficient of humidity, $4.17 \cdot 10^{-7}$ m/s, and the surface heat transfer coefficient, $7 \cdot 10^{-2}$ $W/C \cdot m^2$. Marine environment is implemented using as initial value of free chloride 17.73 kg/m^3, and seasonal variations in exposure using sinewave equations with a range between 10 ºC - 27ºC for temperature and 50% - 60% for humidity. The distance from surface until steel is 0.2 m. This kind of concrete has a relation between water-cement of 0.2.

Figures 1 and 2 show the evolution of free chlorine concentration for different surface mass coefficient of free chlorides. A tenfold reduction in coefficient means that the time for reaching the stable chloride concentration at the metal surface is increased in 35 days.

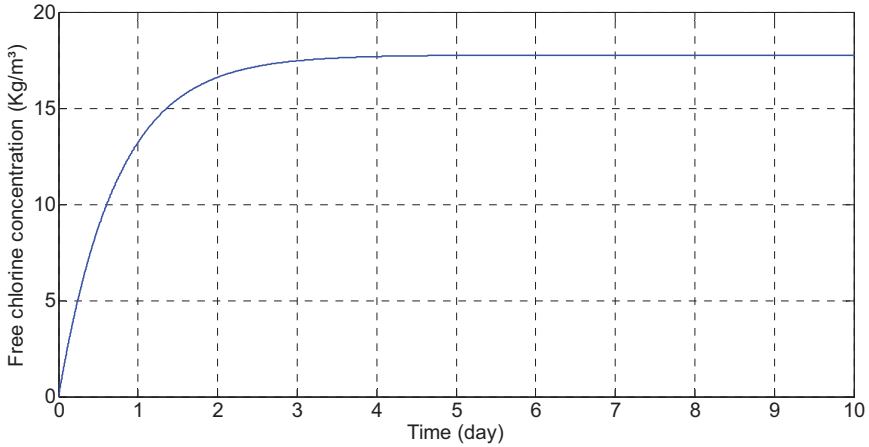


Figure 1 Evolution of free chlorine concentration at steel for a surface mass coefficient of free chlorides of $1 \cdot 10^{-4}$ m/s

 ISBN: 978-84-608-6082-2

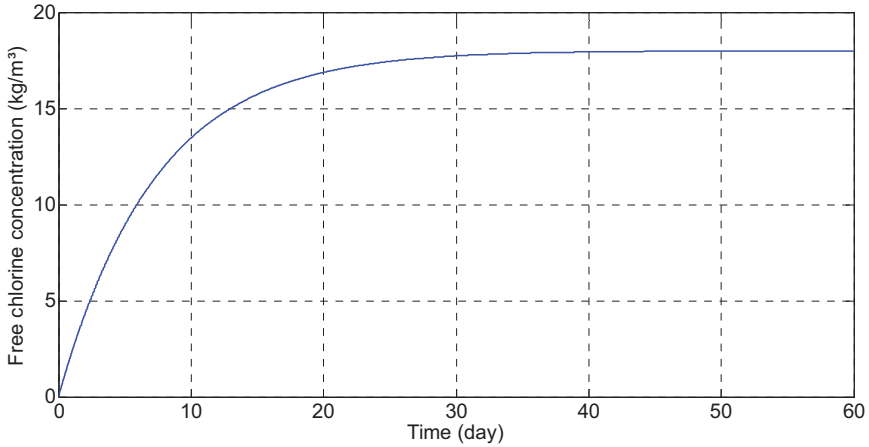


Figure 2 Evolution of free chlorine concentration at steel for a surface mass coefficient of free chlorides of $1 \cdot 10^{-5}$ m/s

5. Conclusions

A numerical model based on network method has been designed and used to successfully simulate the evolution of free chlorine concentration. The influence of the surface mass coefficient of free chlorides is studied. An immediate conclusion can be observed: the progressive slowing down of the progress of chloride penetration when this coefficient diminishes.

6. References

[1] Sanjuán Barbudo, M.A. (2002) Cálculo del período de iniciación de la corrosión de la armadura del hormigón, PhD Thesis, Universidad Complutense De Madrid, Spain.

[2] Martin-Pérez, B., Pantazopoulou, S.J. and Thomas, M.D.A. (2001) Numerical solution of mass transport equations in concrete structures. Computers & Structures, 79 1251-1264.

[3] Saetta, A.V., Scotta, R.V. and Vitaliani R. V. (1993) Analysis of Chloride Diffusion into Partially Saturated Concrete. ACJ Materials Journal. 90:5.

[4] González-Fernández, C.F. and Alhama, F. (2001) Heat Transfer and the Network Simulation Method. J. Horno (ed.) Transworld Research Network, Trivandrum

[5] Microsim Corporation Fairbanks, PSPICE 6.0 (1994) Irvine, California 92718.

[6] NgSpice software [on line]. (quoted 2016) Available on World Wide Web: http://ngspice.sourceforge.net/index.html

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Smoothed particle hydrodynamics method with partially defined fluid particles

Yasutomo Kanetsuki[1], John C. Wells[2] and Susumu Nakata[3]

[1] *Graduate School of Information Science and Engineering, Ritsumeikan University, Japan*
[2] *College of Science and Engineering, Ritsumeikan University, Japan*
[3] *College of Information Science and Engineering, Ritsumeikan University, Japan*

emails: is0061ee@ed.ritsumei.ac.jp, jwells@se.ritsumei.ac.jp, snakata@is.ritsumei.ac.jp

Abstract

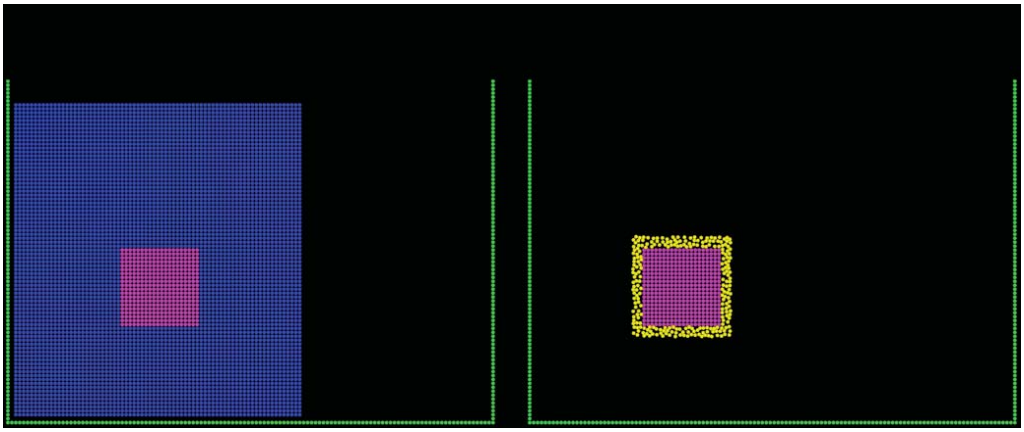
This paper presents a method for nested fluid simulation based on smoothed particle hydrodynamics. Given suitable background flow information of an "external flow" as it evolves in time, our method simulates the motion of particles only within a local material region. In order to perform the simulation, the background physical quantities need to be transferred to the local fluid particles. We employ ghost particles to carry the given physical quantities to the nested fluid. We also solve the problem of density computation appropriately for the ghost particles. Our numerical tests show that accurate local fluid motion can be obtained in such a nested volume of fluid particles.

Key words: smoothed particle hydrodynamics, fluid simulation
MSC2000:

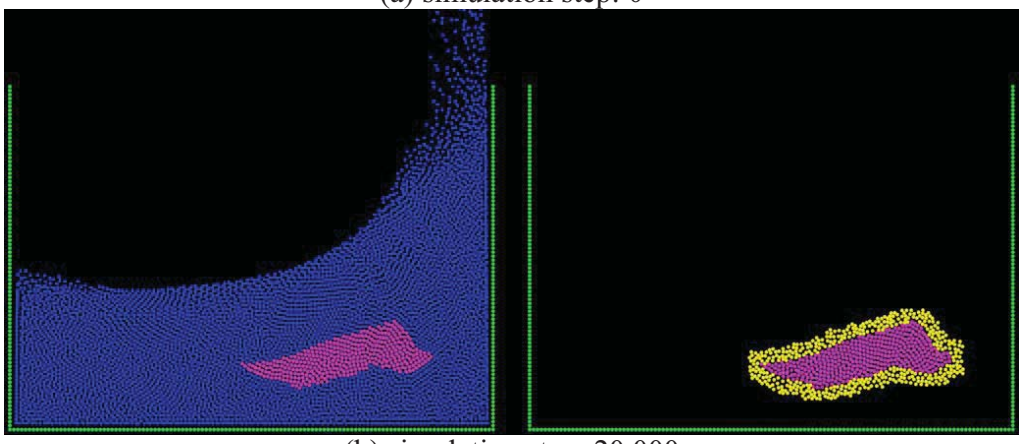
1 Introduction

This paper proposes a method for simulating fluid flow in a nested Lagrangian domain by smoothed particle hydrodynamics (SPH) [1] as exemplified by the labelled region in Figure 1. For geophysical and environmental flows nesting of fine resolution domains into larger domains is the dominant technique to

 ISBN: 978-84-608-6082-2



(a) simulation step: 0



(b) simulation step: 20,000

Fig. 1. Snapshots of particle locations in a typical simulation. The motion of the labelled fluid of fully defined particles (left) and our partially defined particles (right) is compared at the same simulation step.

affordably resolve flow in a local domain that is influenced by complex large-scale phenomena in the surrounding environment.

Our method applies SPH only within a limited material volume and specifies an algorithm to transfer velocity and density from the "outer computation" to the bounding material surface of that volume.

Similarly, the particle tracking method (PTM) [2, 3] tracks fluid particle motion in a given "external" flow field. Generally, the velocities applied in PTM to move particles are simple interpolations that don't satisfy momentum conservation, nor commonly local mass conservation and travelling time. The proposed method considers the physical laws at every time step with SPH, and thereby conserves local mass and momentum within the interpolation limits of SPH.

 ISBN: 978-84-608-6082-2

In particle based simulation of liquid flow, enforcing incompressibility is very important. During initial development, SPH in computer graphics has been successfully applied to simulate liquid flow with interactive speed [4] but doesn't enforce incompressibility. For incompressible fluid, moving particle semi-implicit [5] solves a pressure Poisson equation. Weakly compressible SPH (WCSPH) [6] uses Tait's equation to satisfy a weak compressibility condition while reducing the computational cost. Later, predictive-corrective incompressible SPH [7], implicit incompressible SPH [8] and divergence-free SPH [9] decrease the computational cost by applying a longer time step. This work employs WCSPH for weak incompressibility of fluid flow.

As mentioned above, nested models are typically applied with fixed-grid solvers. For example in [10], a grid based nested model has been developed to offer a cost-effective accurate fluid simulation with high spatial resolution in limited areas. Recently, a hybrid model combining a Boussinesq model and a local SPH model has been also developed to study coastal wave propagation efficiently [11]. In the future, the method proposed here should work to nest a high resolution particle simulation into either a Lagrangian or Eulerian computation of the outer domain. One can anticipate the need for both one-way and two-way coupling, but the current work considers only the simpler one-way coupling.

In the proposed method, the background information needs to be transferred to the partially defined fluid particles. We solve this problem by generating ghost particles around the fluid and interpolating the background information to the ghost particles. We apply Poisson disk sampling (PDS) [12] to distribute the ghost particles. PDS is a technique used for rendering in computer graphics [13] and has been extended to two dimensions [14], arbitrary dimensions [12] and parallel computing [15].

In the originally proposed ghost SPH [16], PDS is used to generate ghost particles and model the effect of air surrounding bodies of liquid. A reduction of computational cost is achieved by avoiding the generation of ghost particles in [17, 18]. In contrast, our ghost particles act to transmit velocity and density from a surrounding liquid.

We found that the standard density formulation [16] combined with ghost particles suffers from serious errors because of the random sampling of positions in PDS. We solve this problem by applying an alternative density formulation that is robust against particle positions. We also simplify PDS to fit fluid simulation with a flag-based method. Our flag-based method avoids the construction of level sets of distance from the fluid surfaces which was used for the ghost particle generation method in [16]. The test results show that our ghost particles

adequately model the effects of the background flow surrounding the local nested fluid region.

2 Nested smoothed particle hydrodynamics

We aim to simulate the motion of fluid within a nested Lagrangian domain, and have adopted SPH for this purpose. We assume that the fluid densities and velocities at the bounding surface of the inner computational domain are available as background information, normally from a computation in the outer domain. Let $\rho^t(\boldsymbol{x})$ be the background fluid density and $\boldsymbol{v}^t(\boldsymbol{x})$ the background velocity defined on the computational domain Ω at an arbitrary time $t \geq 0$. Given a set of local particles with initial positions $\boldsymbol{x}_1^0, \boldsymbol{x}_2^0, \cdots, \boldsymbol{x}_k^0 \in \Omega$, the purpose is to determine their subsequent positions $\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_k^t$ at $t = \Delta t, 2\Delta t, 3\Delta t, \cdots$, where Δt is the time step. Our formulation is based on WCSPH [6], which is designed to compute weakly compressible fluid motion.

In order to perform the simulation, the physical quantities of the background need to be transferred to the nested fluid particles. For this purpose, we employ ghost particles. In this section, we assume that ghost particles fill a layer whose thickness equals the kernel support radius of fluid particles that enclose the nested fluid region at time *t*.

The fluid motion is calculated with WCSPH as follows.

$$\boldsymbol{v}_i^{t+\Delta t} = \boldsymbol{v}_i^t + \Delta t(\boldsymbol{f}_i^{\text{press}} + \boldsymbol{f}_i^{\text{visc}} + \boldsymbol{g})$$
$$\boldsymbol{x}_i^{t+\Delta t} = \boldsymbol{x}_i^t + \Delta t \boldsymbol{v}_i^{t+\Delta t}$$

where $\boldsymbol{v}_i^t$ indicates the velocity of the *i*-th fluid particle at time t, and $\boldsymbol{f}_i^{\text{press}}$, $\boldsymbol{f}_i^{\text{visc}}$ and $\boldsymbol{g}$ respectively denote the accelerations due to pressure, viscous stress, and gravity. In WCSPH, the pressure and viscous accelerations of the *i*-th fluid particle are calculated as follows.

$$\boldsymbol{f}_i^{\text{press}} = -\sum_j m_j \left(\frac{p_i}{\rho_i^2} + \frac{p_j}{\rho_j^2} \right) \nabla_i W_{ij} \tag{1}$$

$$\boldsymbol{f}_i^{\text{visc}} = \begin{cases} \displaystyle\sum_j m_j \frac{2\chi h c_s}{\rho_i + \rho_j} \frac{\boldsymbol{v}_{ij}^{\mathrm{T}} \boldsymbol{x}_{ij}}{\left\| \boldsymbol{x}_{ij} \right\|^2 + 0.01h^2} \nabla_i W_{ij} & (\boldsymbol{v}_{ij}^{\mathrm{T}} \boldsymbol{x}_{ij} < 0) \\ \mathbf{0} & (\text{otherwise}) \end{cases} \tag{2}$$

where h and c_s denote the effective radius and speed of sound in the fluid, m, p, ρ and χ denote the mass, pressure, density, viscosity of each particle, $\boldsymbol{v}_{ij} = \boldsymbol{v}_i - \boldsymbol{v}_j$ and $\boldsymbol{x}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j$, and $W_{ij} = W(\boldsymbol{x}_{ij}, h)$ is the cubic spline kernel

 ISBN: 978-84-608-6082-2

W [19]. The pressure of both the fluid and ghost particles are calculated by Tait's equation [6]:

$$p = \begin{cases} \dfrac{\rho_0 c_s^2}{7}\left(\left(\dfrac{\rho}{\rho_0}\right)^7 - 1\right) & (\rho_0 < \rho), \\ 0 & (\text{otherwise}). \end{cases}$$

In the standard formulation, the density of the i-th fluid particle is computed by

$$\rho_i^t = \sum_j m_j W_{ij} . \tag{3}$$

Note that the index j in (1), (2) and (3) includes both the fluid particles and ghost particles while i doesn't include the ghost particles. Accordingly, the unknown quantities, ρ_j and $\boldsymbol{v}_j$, at the j-th ghost particle need to be determined. For this purpose, we interpolate the given background values to the ghost particle and determine the physical quantities at the nested particles.

In our numerical tests, Equation (3) employed in the original WCSPH causes huge errors because of the random distribution of the position of the ghost particles, i.e., the computed density is too sensitive to the location of the surrounding particles. We found that another formulation,

$$\rho_i^{t+\Delta t} = \rho_i^t + \Delta t \sum_j m_j \boldsymbol{v}_{ij} \cdot \nabla_i W_{ij} , \tag{4}$$

is far less sensitive to the position of the surrounding particles and solves the problem of the instability. Extensions of Equation (4) used in [20, 21] might be considered, but increase computational costs. In this paper, we employ Equation (4) instead of the standard one.

Figure 5 shows one of the results of the density computation in our numerical tests. As can be seen from Figure 5, the density computed by Equation (3) causes serious density fluctuation even a stationary fluid case, but application of Equation (4) duplicates the quantity with sufficient accuracy. Note that the relative density errors (around 0.01%) computed with Equation (4) is sufficiently lower than the compressibility (1%) of our weakly compressible fluid.

Although the reason of the high accuracy of Equation (4) compared to Equation (3) is not proved, the following mechanism relates to the accuracy. Equation (3) depends on only the randomly distributed ghost particle positions but not the physical quantities of the ghost particles. On the other hand, Equation (4) calculates the density depending on the velocities of the ghost particles determined from the background information. The velocity of the ghost particle compensates the error arising from the random ghost particle position. In addition, Equation (4) produces only small variation with small time interval while Equation (3) produces considerably large variation on account of the ghost particle's position varying every time step.

 ISBN: 978-84-608-6082-2

3 Generating ghost particles

Here, we explain how to generate ghost particles around the fluid. In the existing method [16], level sets are applied to compute distance from the fluid surface. We propose a simplified generation method of ghost particles to avoid the construction of the level set.

Instead of the level set values, flags are assigned to each background cell. We then check the validity of particle creation of the sampled ghost particles depending on the flags of the corresponding cell.

The flag in each background cell is assigned to one of the integers 0, 1, 2, 3 and indicates one of the following situations: flag value 0) cell is out of the range of the kernel support radius; flag value 1) cell is within the range of the kernel support radius and far from the fluid surface; flag value 2) cell is within the range of the kernel support radius but near the fluid surface; flag value 3) cell is too close to the fluid particles. The ghost particles are only generated inside cells having a flag value of 1 or 2.

The flag allocation is executed as described in Figure 2. Firstly, generate grids of cells whose edge lengths are $r_{\mathrm{PDS}}/\sqrt{n}$ where r_{PDS} is the sampling radius and n is dimension of the domain (2 or 3), and initialize all flag values to 0. This cell length is specified to allow at most one ghost particle per cell [12]. With initial particle spacing l, we use $r_{\mathrm{PDS}} = 0.8l$ for the sampling radius. Secondly, set flag value 3 if any fluid particles belong to it. Next, set flag value 2 if the value of that cell is not 3 and any part of the cell lies within the sampling radius from any fluid particle. Lastly, set flag value 1 if that flag value is 0 and any part of the cell lies within the kernel support radius αl. For simplicity, we approximate the cells with flag 1 by the cubic (square for two dimensions) shaped cells centered at the cell containing the fluid particle with $2\mathrm{ceil}(\alpha l\sqrt{n}/r_{\mathrm{PDS}})+1$ number of cells for each edge. Here, the function ceil maps a floating-point argument to the least following integer.

After setting the flags, ghost particles are generated by PDS [12]. The sampling starts from the fluid particles in the domain. The ghost particles are generated within the cells having the flag value 1 or 2. The rejection process in PDS is implemented with the neighbor particles not only generated ghost particles but also fluid particles if the sampled position is within the cell with flag value 2. In the case of flag value 1, only the generated ghost particles are applied to the rejection process of PDS. In the computations described herein, we set flags and generate ghost particles at the beginning of every time step.

 ISBN: 978-84-608-6082-2

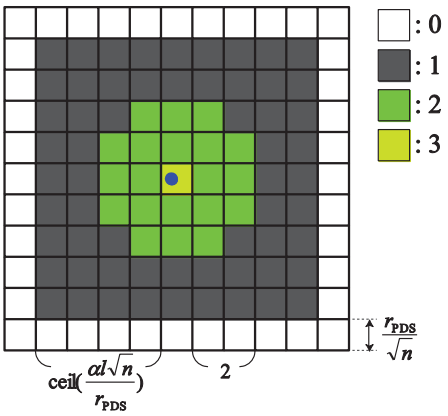


Fig. 2. A flag mapping example for a particle with $\alpha = n = 2$. The point indicates a fluid particle.

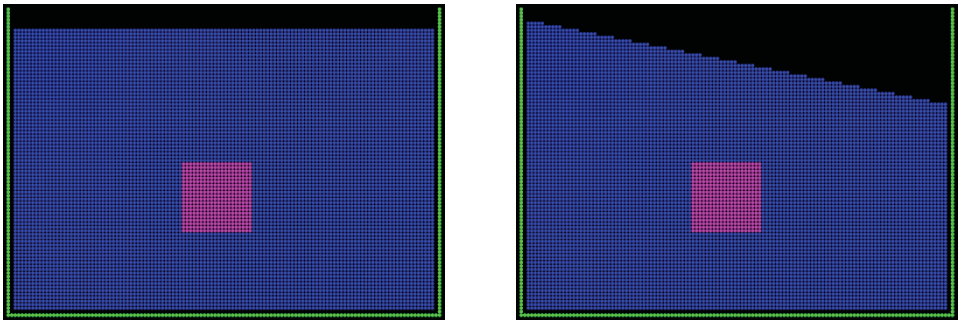
4 Performance analysis

We tested the performance of our method with two kinds of test cases: stationary fluid (Figure 3a) and dynamic fluid (Figure 3b). The tests assess the reproducibility of the fluid motion simulated by the same WCSPH including the density computation for both full simulation in the whole domain and for nested local simulation. We apply the boundary condition used in [6] for both cases. The background flow field is produced by assigning the physical quantities of the nearest neighbor particle gained from the full SPH simulation. In order to reduce the computational cost, we apply counting sort for neighbor particle search as presented in [22] but simulation is implemented on a CPU.

In both test cases, we use the same parameters: $\Delta t = 1.0 \times 10^{-5}$, $m = 4.1086 \times 10^{-4}$, $\rho_0 = 1000$, $\chi = 0.2$, $c_s = 20$, $\alpha = 2$ and $l = \sqrt[n]{m/\rho_0}$. The speed of sound is determined so that the weakly compressible condition (the maximum compressibility is less than 1% for both cases) is satisfied. For the stationary fluid case, we simulate the fluid motion for 200,000 steps in advance to satisfy the stationary condition enough. We compare the obtained physical quantities reproduced by Equation (3) and Equation (4) in the starionary fluid case. We apply the zeroth-order kernel correction [23] for the simulation with Equation (3).

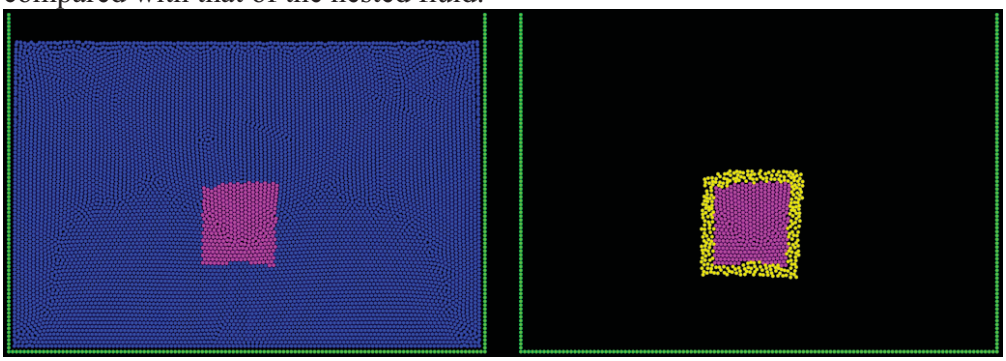
Figure 4 shows the results of full simulation and our local simulation with a stationary fluid case. From Figure 4, no difference can be seen between the full simulation and our local simulation. Figure 5 shows the evolution of the density of the full simulation with Equation (3) (Figure 5a), the local simulation with

 ISBN: 978-84-608-6082-2

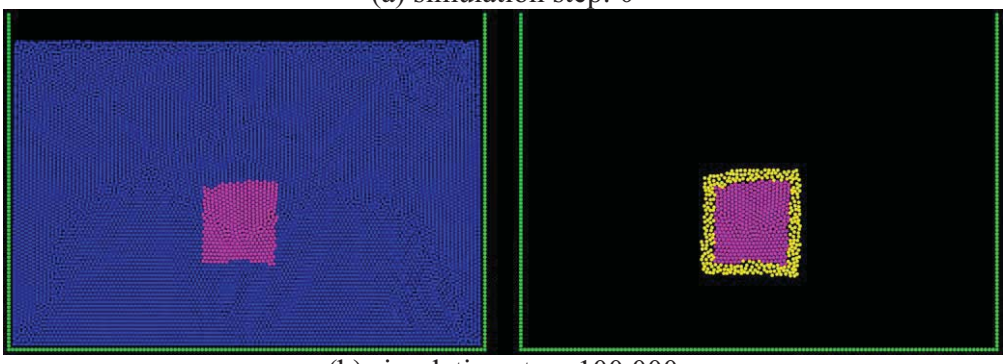


(a) a stationary fluid case (b) a dynamic fluid case

Fig. 3. Our test cases. The motion of the labelled fluid in the full simulation is compared with that of the nested fluid.



(a) simulation step: 0

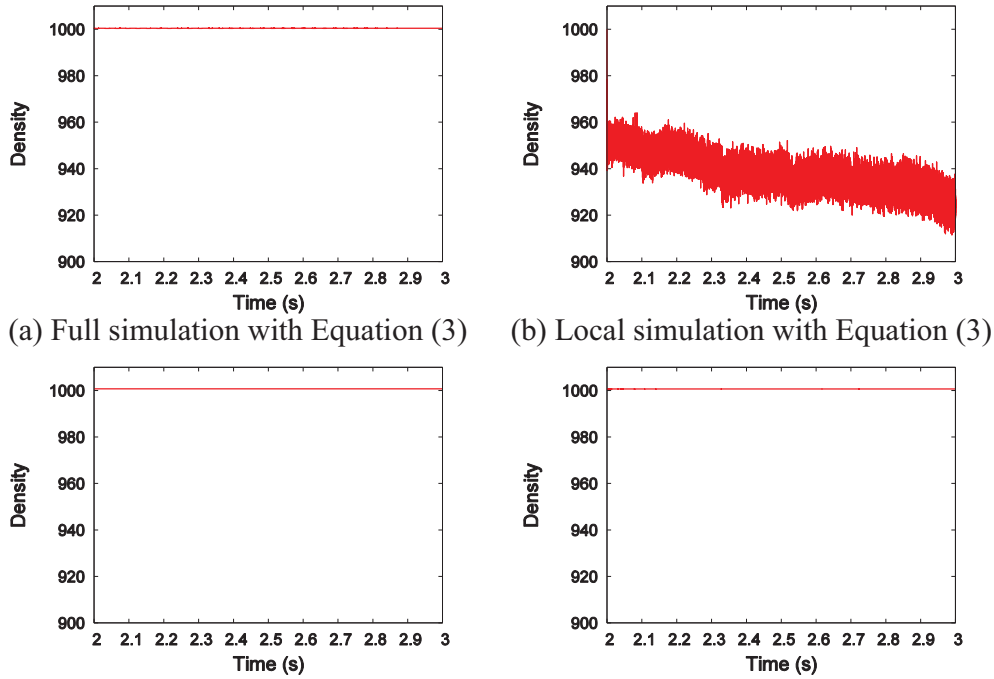


(b) simulation step: 100,000

Fig. 4. The results of a stationary fluid case of (left) the full simulation and (right) our nested local simulation. Motion of the labelled fluid is compared.

Equation (3) (Figure 5b), the full simulation with Equation (4) (Figure 5c) and the local simulation with Equation (4) (Figure 5d). Significant density error can be seen in the local simulation results calculated with Equation (3), but our method with Equation (4) reproduces the physical quantities accurately with a relative density error around 0.01%.

 ISBN: 978-84-608-6082-2



(a) Full simulation with Equation (3) (b) Local simulation with Equation (3)

(c) Full simulation with Equation (4) (d) Local simulation with Equation (4)

Fig. 5. The density evolution in the case of stationary fluid.

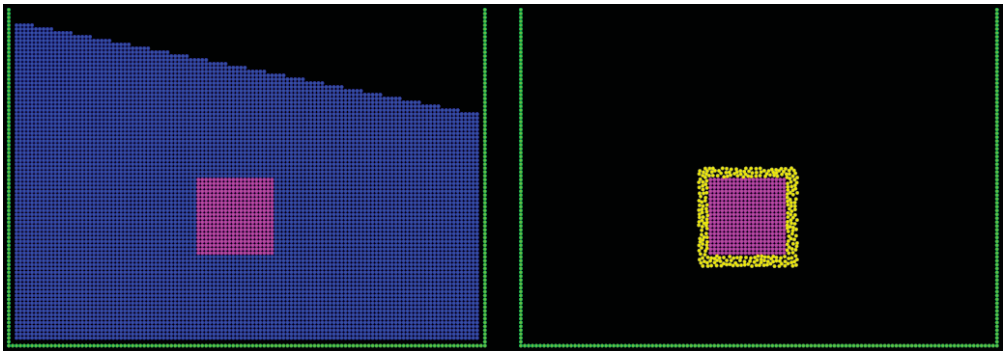
Figure 6 compares the results of full simulation and our local simulation for a dynamic fluid case. From Figure 6, the labelled fluid's motion in the full simulation is perfectly reproduced by our local simulation. Figure 7 shows the evolution of the density of the full simulation and our local simulation. For the density computation, we apply Equation (4) in this dynamic fluid case. The results show that our technique reproduces the fluid quantities accurately even if the fluid moves dynamically.

5 Conclusion

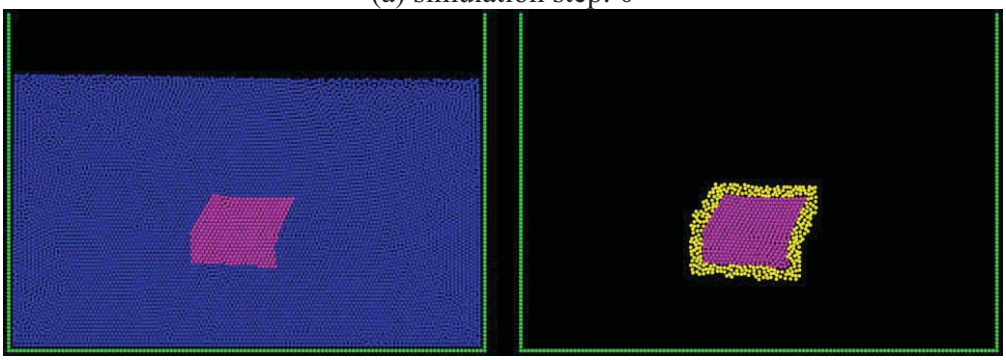
We presented a nested simulation method with locally defined fluid particles. We employ WCSPH to satisfy the weakly compressible condition. The motion of the nested particles follows given background information. In order to transfer the background flow to the fluid, we generate ghost particles around the fluid. We also solve the density computation problem adequately for the ghost particles and simplified the generation of ghost particles with a flag-based method.

The test results show that the proposed technique accurately reproduces the fluid motion of fully simulated fluid. The reproducibility is accurately achieved in both the stationary case and dynamic case.

 ISBN: 978-84-608-6082-2



(a) simulation step: 0



(b) simulation step: 100,000

Fig. 6. The results of a dynamic fluid case of (left) the full simulation and (right) our nested local simulation. Motion of the labelled fluid is compared.

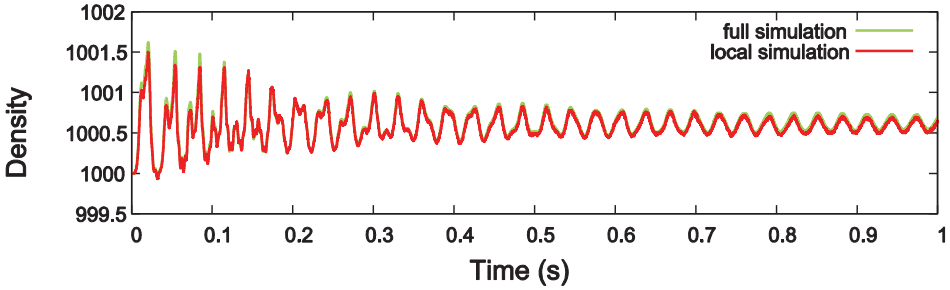


Fig. 7. The density evolution in the case of dynamic fluid with Equation (4).

6 References

[1] J. J. MONAGHAN, *Smoothed particle hydrodynamics*, Ann. Rev. Astron. Astrophys. **30** (1992) 543-574.

[2] D. A. HAM, J. PIETRZAK AND G. S. STELLING, *A streamline tracking algorithm for semi-Lagrangian advection schemes based on the analytic integration of the velocity field*, J. Comp. Appl. Math. **192** (2005) 168-174.

[3] L. POSTMA, J. K. L. VAN BEEK, H. F. P. VAN DEN BOOGAARD AND G. S. STELLING, *Consistent and efficient particle tracking on curvilinear grids*

 ISBN: 978-84-608-6082-2

for environmental problems, Int. J. Numer. Meth. Fluids **71** (2013) 1226-1237.

[4] M. MÜLLER, D. CHARYPAR AND M. GROSS, *Particle-based fluid simulation for interactive applications*, ACM SIGGRAPH/Eurographics Symp. on Comp. Animat. (2003) 154-159.

[5] S. KOSHIZUKA AND Y. OKA, *Moving-particle semi-implicit method for fragmentation of incompressible fluid*, Nucl. Sci. Eng. **123** (1996) 421-434.

[6] M. BECKER AND M. TESCHNER, *Weakly compressible SPH for free surface flows*, ACM SIGGRAPH/Eurographics Symp. on Comp. Animat. (2007) 209-217.

[7] B. SOLENTHALER AND R. PAJAROLA, *Predictive-corrective incompressible SPH*, ACM Trans. Graph. **28** (2009) Article No. 40.

[8] M. IHMSEN, J. CORNELIS B. SOLENTHALER, C. HORVATH AND M. TESCHNER, *Implicit incompressible SPH*, IEEE Trans. Vis. Comput. Graph. **20** (2014) 426-435.

[9] J. BENDER AND D. KOSCHIER, *Divergence-free smoothed particle hydrodynamics*, ACM SIGGRAPH/Eurographics Symp. on Comp. Animat. (2015) 147-155.

[10] S. NASH AND M. HARTNETT, *Development of a nested coastal circulation model: boundary error reduction*, Environ. Modell. Softw. **53** (2014) 65-80.

[11] M. NARAYANASWAMY, A. J. C. CRESPO, M. GÓMEZ-GESTEIRA AND R. A. DALRYMPLE, *SPHysics-FUNWAVE hybrid model for coastal wave propagation*, J. Hydraul. Res. **48** (2010) 85-93.

[12] R. BRIDSON, *Fast Poisson disk sampling in arbitrary dimensions*, ACM SIGGRAPH 2007 sketches (2007) Article No. 22.

[13] R. L. COOK, *Stochastic sampling in computer graphics*, ACM Trans. Graph. **5** (1986) 51-72.

[14] D. DUNBAR AND G. HUMPHREYS, *A spatial data structure for fast Poisson-disk sample generation*, ACM Trans. Graph. **25** (2006) 503-508.

[15] L.-Y. WEI, *Parallel Poisson disk sampling*, ACM Trans. Graph. **27** (2008) Article No. 20.

[16] H. SCHECHTER AND R. BRIDSON, *Ghost SPH for animating water*, ACM Trans. Graph. **31** (2012) Article No. 61.

[17] N. AKINCI, G. AKINCI AND M. TESCHNER, *Versatile surface tension and adhesion for SPH fluids*, ACM Trans. Graph. **32** (2013) Article No. 182.

[18] X. HE, H. WANG, F. ZHANG, H. WANG, G. WANG AND K. ZHOU, *Robust simulation of small-scale thin features in SPH-based free surface flows*, ACM Trans. Graph. **34** (2014) Article No. 7.

[19] J. J. MONAGHAN, *Smoothed particle hydrodynamics*, Rep. Prog. Phys. **68** (2005) 1703-1759.

 ISBN: 978-84-608-6082-2

[20]M. ANTUONO, A. COLAGROSSI, S. MARRONE AND D. MOLTENI, *Free-surface flows solved by means of SPH schemes with numerical diffusive terms*, Comput. Phys. Commun. **181** (2010) 532-549.
[21]S. MARRONE, M. ANTUONO, A. COLAGROSSI, G. COLICCHIO, D. LE TOUZÉ AND G. GRAZIANI, *δ-SPH model for simulating violent impact flows*, Comput. Method. Appl. M. **200** (2011) 1526-1542.
[22]R. C. HOETZLEIN, *Fast fixed-radius nearest neighbors: interactive million-particle fluids*, GPU Technol. Conf. (2014).
[23]J. BONET AND S. KULASEGARAM, *A simplified approach to enhance the performance of smoothed particle hydrodynamics methods*, Appl. Math. Comput. **126** (2002) 133-155.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Density Matrix Simulations of Quantum Electron Dynamics in Perturbed Atoms and Electron Gas

Hikaru Kitamura
Department of Physics, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

email: kitamura@scphys.kyoto-u.ac.jp

Abstract

Equations of motion for density matrices are formulated for interacting electron systems by taking into account static Hartree-Fock energy levels and wave functions as input. The resultant quantum kinetic equations can simulate time evolutions of excitation and relaxation of electrons after time-varying external fields are switched on. Applications to ionizing atoms and bulk electron gas are presented.

Key words: density matrix, quantum electron dynamics, orbital relaxation, electron gas

1. Introduction

In quantum mechanics, the motion of particles may generally be analyzed from a viewpoint of transitions among quantum states. When an external field is applied to an interacting many-electron system, for instance, some electrons are excited from occupied to unoccupied orbitals, leaving holes behind. The holes in turn exert attractive Coulomb forces on the other electrons, modifying their spatial distributions and energy levels. We have been developing *quantum kinetic equations* with the density matrix formalism based on such a viewpoint [1].

Our theory takes into account unrestricted Hartree-Fock (UHF) calculations of energy levels and wave functions in the initial unperturbed state as input; we then introduce density matrices represented in terms of the those UHF orbitals. After an external field is switched on, the dynamics of electrons and holes are simulated through time evolutions of those density matrices without computing wave functions any more.

 ISBN: 978-84-608-6082-2

The structure of the density matrix equations is highly nonlinear so that the validity of their numerical solutions should be checked carefully through comparison with analytical results. In the CMMSE14 Conference, we have examined elementary problems such as the response of noninteracting electrons to a locally applied potential well [2]. In this work, numerical analysis is extended to *interacting* electrons; orbital relaxation dynamics in an ionizing atom is reported in Sec. 3.1. The theory is also applicable to the dynamics, transport, and impurity problems in bulk electron gas, as we shall briefly summarize in Sec. 3.2.

2. Basic Formalism

Let us consider a system of N_α spin-up and N_β spin-down electrons in the potential field of fixed ions. To begin with, we solve the UHF equation [3] self-consistently for the initial (t = 0) ground state, and compute the energy levels $\varepsilon_{k\sigma}$ and wave functions $\psi_{k\sigma}(\mathbf{r})$ for a set of single-particle states k (=1,2,$\cdots$,N_{AO}) with spin σ (=α, β). If the usual basis-set expansion technique [3] is used here, the total number of atomic orbitals so obtained (N_{AO}) is limited to the size of the basis set.

Suppose that an external potential field $v_{\mathrm{ext}}(\mathbf{r},\sigma,t)$ is applied to this system for time $t > 0$. To compute subsequent electron dynamics, the matrix elements

$$\left(v_{\mathrm{ext}}(t)\right)_{kk'\sigma} \equiv \int d\mathbf{r}\psi^*_{k\sigma}(\mathbf{r})v_{\mathrm{ext}}(\mathbf{r},\sigma,t)\psi_{k'\sigma}(\mathbf{r}) \tag{1}$$

and the two-electron Coulomb repulsion integrals

$$V^{\sigma\sigma'}_{k_1k_2k_3k_4} \equiv \int d\mathbf{r}_1 \int d\mathbf{r}_2 \psi^*_{k_1\sigma}(\mathbf{r}_1)\psi_{k_2\sigma}(\mathbf{r}_1)\frac{e^2}{|\mathbf{r}_1-\mathbf{r}_2|}\psi^*_{k_3\sigma'}(\mathbf{r}_2)\psi_{k_4\sigma'}(\mathbf{r}_2) \tag{2}$$

should be evaluated in advance. We note here that the indices k, k', etc., extend from occupied ($1 \le k \le N_{\alpha,\beta}$) to virtual ($N_{\alpha,\beta}+1 \le k \le N_{\mathrm{AO}}$) orbitals; the mixing of those orbitals plays an essential role in orbital relaxation illustrated later in Sec. 3.1.

The dynamics of electrons obeys the time-dependent Schrödinger equation $i\hbar\partial\Psi(t)/\partial t = H(t)\Psi(t)$ for the total wave function $\Psi(t)$. Instead of $\Psi(t)$, we simulate the one-electron density matrix defined as $\langle\rho_{kk'\sigma}(t)\rangle \equiv \left\langle\Psi(t)\left|c^\dagger_{k\sigma}c_{k'\sigma}\right|\Psi(t)\right\rangle$, which in turn depends on the two-electron density matrix $\langle\rho^{\sigma\sigma'}_{ijkl}(t)\rangle \equiv \left\langle\Psi(t)\left|c^\dagger_{i\sigma}c^\dagger_{j\sigma'}c_{l\sigma'}c_{k\sigma}\right|\Psi(t)\right\rangle$ [4]. Here, $c^\dagger_{k\sigma}$ and $c_{k\sigma}$ represent the creation and annihilation operators, respectively, for state $\{k,\sigma\}$.

 ISBN: 978-84-608-6082-2

Since $\left\langle \rho_{ijkl}^{\sigma\sigma'}(t)\right\rangle = \left\langle \rho_{ik\sigma}(t)\right\rangle\left\langle \rho_{jl\sigma'}(t)\right\rangle - \delta_{\sigma\sigma'}\left\langle \rho_{jk\sigma}(t)\right\rangle\left\langle \rho_{il\sigma}(t)\right\rangle$ in UHF approximation [4], the time-dependent Schrödinger equation can be reduced [1] to coupled nonlinear differential equations for $\left\langle \rho_{kk'\sigma}(t)\right\rangle$ that involve matrix elements (1) and (2). Here, the attractive electron-hole interaction enters through the *self-energy matrix*

$$\tilde{\varepsilon}_{kk'\sigma}(t) \equiv \sum_{k_1k_2\sigma_1}\left(V_{kk'k_1k_2}^{\sigma\sigma_1} - \delta_{\sigma\sigma_1}V_{kk_2k_1k'}^{\sigma\sigma}\right)\left[\left\langle \rho_{k_1k_2\sigma_1}(t)\right\rangle - \delta_{k_1k_2}f_{k_1\sigma_1}(0)\right]. \qquad (3)$$

It is convenient to decompose the density matrix equations further into the diagonal part $f_{k\sigma}(t) \equiv \left\langle \rho_{kk\sigma}(t)\right\rangle$ representing the occupation number of state $\{k, \sigma\}$ and the off-diagonal part $\left\langle \rho_{kk'\sigma}(t)\right\rangle$ related to the configuration mixing between states k and k'. The final expressions for the density matrix equations can be found in Ref. [1].

Initially, the density matrix is diagonal and hence $\left\langle \rho_{kk'\sigma}(0)\right\rangle = 0$ for $k \neq k'$; it then follows from (3) that $\tilde{\varepsilon}_{kk'\sigma}(0) = 0$. The initial diagonal elements are set as $f_{k\sigma}(0) = 1$ for $k \leq N_\sigma$ and $f_{k\sigma}(0) = 0$ for $N_\sigma + 1 \leq k \leq N_{\mathrm{AO}}$.

3. Applications

3.1. Atomic orbital relaxation

Orbital relaxation associated with core-level ionization of an atom or a bulk material plays a crucial role in the interpretations of x-ray absorption and emission spectra [5]. Density matrix simulations of orbital relaxation dynamics have been carried out for *K*-shell ionization of a Na atom (N_α = 6, N_β = 5). Initial UHF calculations have been performed by using the Slater-type 7s6p basis set (N_{AO} = 25), as shown by the left panel of Fig. 1. The 1s electron has been removed gradually for $t > 0$ by introducing a phenomenological ionization term $-w_{\mathrm{ion}}[1-\exp(-t/\tau)]f_{1s,\alpha}(t)$ in the equation for $\partial f_{1s,\alpha}(t)/\partial t$, with an ionization rate w_{ion} and a time constant τ for an onset of the perturbation. The 1s hole so created produces an attractive potential field, to which the outer electrons respond and adjust themselves. The ionization term is turned off when the total number of electrons $N(t) \equiv \sum_{k\sigma} f_{k\sigma}(t)$ is decreased exactly by 1 and hence the atom becomes a singly charged ion. The middle panel of Fig. 1 depicts time evolutions of the energy levels evaluated as the eigenvalues of the instantaneous Fock matrix

$$F_{kk'\sigma}(t) = \varepsilon_{k\sigma}\delta_{kk'} + \tilde{\varepsilon}_{kk'\sigma}(t). \qquad (4)$$

 ISBN: 978-84-608-6082-2

For a comparison, the right panel exhibits the energy levels of a Na^+ ion with a core hole in the 1sα orbital, obtained by a separate UHF calculation with the same basis set as above.

We assume that the ejected electron does not interact with the remaining electrons. Then, provided that $w_{\rm ion}$ is small and τ is long enough, the *adiabatic theorem* ensures that the simulation should automatically converge to the ground state of a singly charged ion with *fully relaxed* orbitals. This conjecture is corroborated in Fig. 1, where we find that each energy level displayed in the middle panel eventually approaches the corresponding eigenstate of an ion plotted in the right panel.

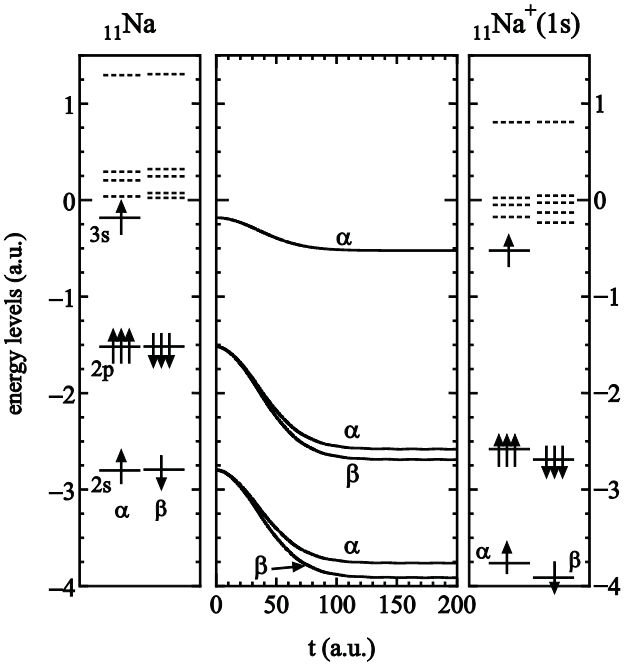


FIGURE 1: Orbital relaxation dynamics in the *K*-shell ionization of a Na atom. The left panel depicts the UHF energy levels $\varepsilon_{k\sigma}$ in the ground state. The solid and dotted lines correspond to occupied and unoccupied orbitals, respectively. The 1s levels (not shown) are located at $\varepsilon_{1s,\sigma} = -40.48$ a.u. The middle panel shows eigenvalues of the instantaneous Fock matrix (4) computed through numerical solutions to the density matrix equations. The simulation parameters are $w_{\rm ion} = 0.1$ a.u. and $\tau = 100$ a.u. The right panel is same as the left panel but for a Na^+ ion with a core hole in 1sα orbital.

 ISBN: 978-84-608-6082-2

In orbital relaxation, not only the energy levels but the orbital shapes are modified. The latter may be caused by a configuration mixing of state k with other states k', which is detectable through an off-diagonal matrix element $\langle \rho_{kk'\sigma}(t) \rangle$. Numerical results reveal that some of the matrix elements $\langle \rho_{kk'\sigma}(t) \rangle$ for occupied (k) and virtual (k') orbitals take on nonnegligible values after convergence ($t \approx 200$ a.u.); this means that the *Koopmans' theorem* on the ionization energy $I_{\mathrm{p}} \approx -\varepsilon_{1s\alpha}$ [3], which neglects orbital relaxation, is no longer satisfied due to a strong perturbation by the core hole.

3.2. Other applications

By appropriately choosing the onset time scale of external fields τ and the number of electrons N, the density matrix equations can demonstrate quantum electron dynamics in various settings from small to bulk systems and from adiabatic to nonadiabatic perturbation regime, as sketched in Fig. 2. Brief summary of recent progress is given below.

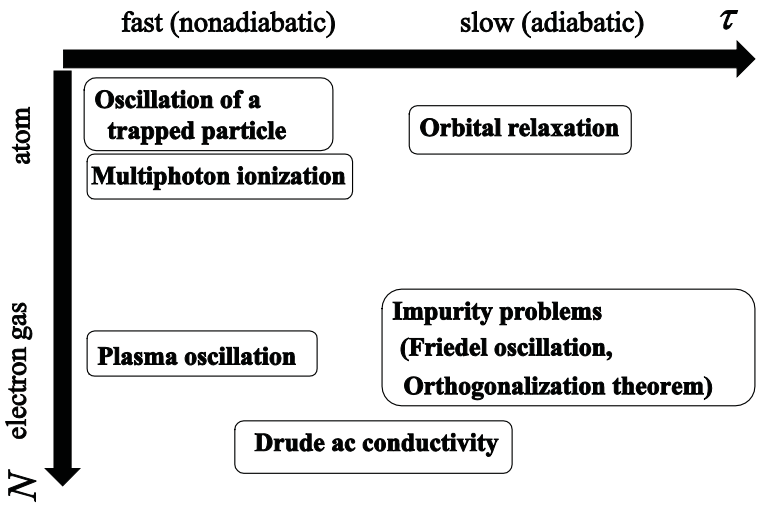


FIGURE 2: Overview of physical phenomena studied with the density matrix equations for various combinations of τ and N.

(a) *Oscillation of a trapped particle*
Consider a particle confined in a large box. When an attractive potential well is applied suddenly near the centre of the box, the quantum distribution of the particle exhibits oscillation around the potential well, whose frequency is

determined by the energy difference between the ground and first excited states [1,2]. The physical origin of the oscillation can be attributed to an electronic excitation across the Landau-Zener type pseudo-crossing of those two potential energy curves.

(b) *Multiphoton ionization*
Under an intense laser field, a bound electron in an atom can be excited to continuum via absorption of multiple photons. According to the pioneering theory of Keldysh [6], the kinetic energy ε of the photoelectron emitted through n-photon absorption is given by $\varepsilon = n\hbar\omega - I_{\mathrm{p}} - U_{\mathrm{p}}$, where ω is the photon frequency, I_{p} is the ionization energy of the bound electron; $U_{\mathrm{p}} \equiv e^2 E_0^2 / 4m\omega^2$ is the ponderomotive energy, defined as an average kinetic energy of a free electron of charge $-e$ and mass m under an oscillating electric field $E_0\cos(\omega t)$. We have proven [1] that the Keldysh formula for the multiphoton ionization rate of a hydrogen atom can be reproduced analytically from the density matrix equations through approximating the continuum wave function by a plane wave. Numerical analysis of the multiphoton ionization has also been made with the aid of the spherical-harmonics expansion of the continuum density matrices [7].

(c) *Plasma oscillation*
Plasma oscillation is a long-wavelength collective motion of charged particles triggered by a local charge imbalance [8]. The relation between the frequency ω and the wave vector $\boldsymbol{q}$ of the plasma oscillation depends on the Planck constant, reflecting its quantum-mechanical nature. For an electron gas, whose single-particle wave function $\psi_{\mathbf{k}\sigma}(\mathbf{r})$ is given by the plane wave $\exp(i\mathbf{k}\cdot\mathbf{r})$, we have demonstrated plasma oscillations in q-t space by solving the density matrix equations numerically using the spherical-harmonics expansion method [7]. Here, the diagonal matrix elements $f_{k\sigma}(t)$ remain constant whereas the off-diagonal elements $\left\langle \rho_{\mathbf{k},\mathbf{k}+\mathbf{q}}(t) \right\rangle$ exhibit oscillations driven by the self-energy matrix $\tilde{\varepsilon}_{\mathbf{k}+\mathbf{q},\mathbf{k}}(t)$. When the exchange term involving $\delta_{\sigma\sigma_1}$ in (3) is neglected, the theory becomes equivalent to the conventional random-phase approximation [8].

(d) *Impurity problems*
When a charged impurity is inserted into an electron gas, the tail of the induced electron density exhibits a spatial oscillation called *Friedel oscillation* [1,2]; it originates from a discontinuity in the Fermi distribution function across the Fermi level in the unperturbed state. For this system, Anderson also predicted the *orthogonalization theorem*, which states that the overlap integral of the Slater determinants with and without the impurity vanishes in the limit of $N \to \infty$; it is a consequence of catastrophic low-energy electronic excitations across the Fermi

level [1,2]. These are both static phenomena but can be reproduced by the density matrix equations for uniform electron gas to which the impurity potential is slowly (adiabatically) applied [1,2].

(e) *Drude formula of ac conductivity*
The Drude formula, proposed more than one hundred years ago, is well established as a standard expression for the electric conductivity due to itinerant noninteracting electrons in metals and semiconductors [9]. For an ac electric field with frequency ω, it reads

$$\tilde{\sigma}(\omega) = \frac{n_{\mathrm{e}} e^2 \tau_{\mathrm{coll}}}{m_*} \frac{1}{1 - i\omega\tau_{\mathrm{coll}}}, \tag{5}$$

where m_* and n_{e} represent the effective mass and the number density of the electrons, respectively; τ_{coll} refers to the collisional relaxation time. In most textbooks (e.g., [9]), this formula is derived either from the classical Newton's equation of motion or the semiclassical Boltzmann equation. Although the Kubo formula offers a general quantum-mechanical expression of conductivities [10], it has been a long-standing difficulty to derive (5) from Kubo formula especially for low-frequency regime $\omega\tau_{\mathrm{coll}} < 1$, due in part to a complexity in the treatment of phonon-assisted intraband transitions [11]. Alternatively, we have shown [12] that linear-response analysis of the density matrix equations, which avoids a perturbative treatment of electron-phonon scattering, facilitates a lucid quantum-mechanical derivation of formula (5) over the entire ω-regime.

4. References

[1] H. KITAMURA, *Density-matrix theory of quantum dynamics under a strong external field switched on nonadiabatically*, Int. J. Quant. Chem. **114** (2014) 1518-1527.

[2] H. KITAMURA, *Quantum oscillation phenomena induced by strong nonadiabatic perturbation*, Proceedings of the 2014 International Conference on Computational and Mathematical Methods in Science and Engineering, Vol. V, edited by J. Vigo-Aguiar (2014) 1350-1358.

[3] A. SZABO AND N.S. OSTLUND, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Macmillan, 1982.

[4] J. CIOSLOWSKI, ED., *Many-Electron Densities and Reduced Density Matrices*, Kluwer, New York, 2000.

[5] T. PRIVALOV, F. GEL'MUKHANOV AND H. ÅGREN, *Role of relaxation and time-dependent formation of x-ray spectra*, Phys. Rev. B **64** (2001) 165115.

[6] L.V. KELDYSH, *Ionization in the field of a strong electromagnetic wave*, Sov. Phys. JETP **20** (1965), 1307-1314.

 ISBN: 978-84-608-6082-2

[7] H. KITAMURA, *Spherical-Harmonics Expansion Method for Density-Matrix Simulations of Quantum Electron Dynamics in Continuum States*, Int. J. Quant. Chem. **115** (2015) 1587-1596.
[8] S. ICHIMARU, *Statistical Plasma Physics, Vol. I: Basic Principles*, Addison-Wesley, Reading, MS, 1992.
[9] N.W. ASHCROFT AND N.D. MERMIN, *Solid State Physics*, Brooks/Cole, 1976.
[10] R. KUBO, *Statistical-Mechanical Theory of Irreversible Processes I. General Theory and Simple Applications to Magnetic and Conduction Problems*, J. Phys. Soc. Japan **12** (1957) 570-586.
[11] G. DE FILIPPIS, V. CATAUDELLA, A. DE CANDIA, A.S. MISHCHENKO AND N. NAGAOSA, *Alternative representation of the Kubo formula for optical conductivity: A shortcut to transport properties*, Phys. Rev. B **90** (2014) 014310.
[12] H. KITAMURA, *Derivation of the Drude conductivity from quantum kinetic equations*, Eur. J. Phys. **36** (2015) 065010.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Solving the Schrodinger Equation of Harmonium Systems with the Free-Complement Local-Schrödinger-Equation method

Yusaku I. Kurokawa[1] and Hiroshi Nakatsuji[1]

[1] *Quantum Chemistry Research Institute,*
Kyodai Katsura Venture Plaza, North building 107,
1-36 Goryo-Oohara, Nishikyo-ku, Kyoto, 615-8245 Japan

emails: y.kurokawa@qcri.or.jp, h.nakatsuji@qcri.or.jp

Abstract

We solved the Schrödinger equation of the two- and three-electron harmonium systems with the Free Complement - Local Schrödinger Equation (FC-LSE) method. The results agreed with the Cioslowski's energies to five digits, and, in some cases, a little lower in the sixth digits. It was observed that the r_{ij} terms are important in both the two- and three-electron harmonium systems.

Key words: Free Complement method, Harmonium

1. Introduction

A harmonium system is an artificial and imaginary system but important because it is an exactly solvable system [1]: it is also a good model of a quantum dot [2]. Harmonium atoms consist of a nucleus and some electrons, where the electron-nucleus potentials are harmonic and the electron-electron potentials are Coulombic. The Hamiltonian of the harmonium is written as

$$H = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \frac{\omega^2}{2}\sum_{i=1}^{N} r_i^2 + \sum_{j>i=1}^{N}\frac{1}{r_{ij}}, \qquad (1)$$

where N is the number of electrons, and ω is a force constant. The two-electron harmonium (N = 2) was first studied by Kestner and Sinanoğlu [1], and they found that the exact wave function is expressed in a closed form with a r_{12} term. The two- and three- electron harmoniums are widely studies by Cioslowski, Taut, Pernal, Staemmler, Strasburger, Matito, Trickey and Loos *et al*. [3-8].

 ISBN: 978-84-608-6082-2

When the last term (Coulombic term) in Eq. (1) does not exist, the exact wave function of this system in the ground state can be written as

$$\psi = A\left[\exp\left(-a_1 r_1^2\right)\exp\left(-a_2 r_2^2\right)\theta\right] \tag{2}$$

for $N = 2$, and

$$\psi = A\left[\exp\left(-a_1 r_1^2\right)\exp\left(-a_2 r_2^2\right) z_3 \exp\left(-a_3 r_3^2\right)\theta\right] \tag{3}$$

for $N = 3$, where A is the anti-symmetrizer , $a_1 = a_2 = a_3 = \omega/2$, and θ is a proper sin function.

In the two-electron harmonium, where there exists the Coulombic term, the exact wave function in the ground state is written as

$$\psi = \exp\left(-\frac{1}{4}r_1^2\right)\exp\left(-\frac{1}{4}r_2^2\right)\left(1+\frac{1}{2}r_{12}\right)(\alpha\beta - \beta\alpha) \tag{4}$$

when $\omega = 1/2$ [2] . The $(1 + r_{12}/2)$ term in Eq. (4) is additionally multiplied to the non-Coulombic-term system (Eq. (2)) due to the Coulombic terms in the Hamiltonian. In the three-electron harmonium, unfortunately, the exact wave function is not found in a closed form yet.

Solving the Schrödinger equation (S.E.) of any atoms and molecules is one of the goals for quantum chemists because it governs the world of chemistry. In 2004, one of the authors proposed the Free Complement (FC) method, with which we can solve the Schrödinger equations exactly [9]. When analytical integration is not possible, we use the FC-Local Schrödinger Equation (FC-LSE) method [10-13].

The FC-LSE method is based on a sampling method, which can be applied to any systems without integration problems. We have reported some applications of the FC-LSE method to some organic and inorganic molecules [12, 14].

In this presentation, we apply the FC-LSE method to solve the Schrödinger equations of the two- and three-electron harmonium systems.

2. Free-Complement Local-Schrödinger-Equation method

The details of the FC-LSE method were explained in Ref. [11-13]. The FC wave function is generated as follows, using the simplest iterative complement (SIC) formula given by

$$\psi^{(n)} \equiv \left[1 + C_{n-1} g\left(H - E_{n-1}\right)\right]\psi^{(n-1)} , \tag{5}$$

where H is the Hamiltonian of the system (Eq. (1)). The g function was introduced to cancel the divergence in the Hamiltonian: For the two-electron harmonium, we take

$$g = r_{12} , \tag{6}$$

and for the three-electron harmonium

$$g = \sum_{j>i=1}^{3} r_{ij} . \tag{7}$$

 ISBN: 978-84-608-6082-2

It was proven that the SIC formula leads to the exact wave function at convergence. The initial wave function $\psi^{(0)}$ is arbitrary if it has an overlap with the exact wave function. Operating *gH* and *g* operators on $\psi^{(0)}$ *n* times according to Eq. (5), we can pick up linearly independent analytical functions $\{\phi^{(n)}\}$. The FC wave function at order *n* is expressed in a linear combination as

$$\psi^{(n)} = \sum_{i=1}^{M^{(n)}} C_i^{(n)} \phi_i^{(n)}, \tag{8}$$

where $\{\phi^{(n)}\}$ is referred to *complement functions* (cfs) and $M^{(n)}$ is the number of the cfs. The $\psi^{(n)}$ determined by Eq. (8) converges to the exact solution of the SE as the order *n* increases, if the coefficients $\{C^{(n)}\}$ are correctly determined. In the FC-LSE method, the coefficients $\{C^{(n)}\}$ are determined as follows: When ψ satisfies the Schrödinger Equation, the following equation holds for any point $\mathbf{r}$,

$$H\psi(\mathbf{r}) = E\psi(\mathbf{r}). \tag{9}$$

Inserting $\psi^{(n)}$ into Eq. (9), we obtaine

$$\sum_{i=1}^{M^{(n)}} C_i^{(n)} H\phi_i^{(n)}(\mathbf{r}_\mu) = E\sum_{i=1}^{M^{(n)}} C_i^{(n)} \phi_i^{(n)}(\mathbf{r}_\mu) \quad (\mu = 1..N_s), \tag{10}$$

for any sampling points $\{\mathbf{r}_\mu\}$ where N_s is the number of the sampling points. We use Eq. (10) to determine the coefficients $\{C^{(n)}\}$ and energy *E*. Eq. (10) can be written in matrix form as $\mathbf{AC} = E\mathbf{BC}$, where $A_{\mu i} = H\phi_i^{(n)}(\mathbf{r}_\mu)$ and $B_{\mu i} = \phi_i^{(n)}(\mathbf{r}_\mu)$. Multiplying $\mathbf{B}^\dagger$ from the left, we get $\mathbf{HC} = E\mathbf{SC}$ where $\mathbf{H} = \mathbf{B}^\dagger\mathbf{A}$ and $\mathbf{S} = \mathbf{B}^\dagger\mathbf{B}$, and the (*i*,*j*)-th elements of $M^{(n)}$ x $M^{(n)}$ matrices **H** and **S** are calculated by $H_{ij} = \sum_{\mu=1}^{N_s} \phi_i(\mathbf{r}_\mu) \bullet H\phi_j(\mathbf{r}_\mu)$ and $S_{ij} = \sum_{\mu=1}^{N_s} \phi_i(\mathbf{r}_\mu) \bullet \phi_j(\mathbf{r}_\mu)$, respectively. By solving Eq. $\mathbf{HC} = E\mathbf{SC}$, we obtain the total energy *E* and corresponding coefficients $\{C^{(n)}\}$.

3. Two-electron harmonium

First we calculate the two-electron harmonium with the FC-LSE method [13]. The Hamiltonian is given in Eq. (1) and we set ω = 1/2. In this case the exact solution is known in a closed form [2].

The initial function $\phi^{(0)}$ we employed is Eq. (2) with $\theta = (\alpha_1\beta_2 - \beta_1\alpha_2)$, since it is the exact wave functions of the non-Coulombic term system. At order *n* = 1, the generated cfs are $\{\phi\} = \{\phi^{(0)}, r_{12}\phi^{(0)}, (r_1^2 + r_2^2)\phi^{(0)}\}$. To determine their linear combination coefficients, we need at least two sampling points if we ignore the normalization constant. The position of the sampling points is arbitrary in this case since the generated cfs are complete. For example, using only two sampling points $\{\mathbf{r}_\mu\} = \{(x_1, y_1, z_1, x_2, y_2, z_2)\}$ = {(-1.1,0.3,0.4,-0.5,1.2,-2.3), (1.2,0.2,-1.2,0.8,2.5,-0.9)}

 ISBN: 978-84-608-6082-2

$(\mu = 1,2)$ in Eq. (10) gives $\{C\} = \{1,1/2,0\}$ and E = 2 au exactly. This is the exact solution of the two-electron harmonium [2]. Even if we use any other two sampling points, we obtain the same exact results. At the higher order *n*, different cfs are generated, but their coefficients become exactly zero with arbitrary sampling points.

4. Three-electron harmonium

Next we study the three-electron harmonium. The force constant ω in Eq. (1) is taken to be ω = 10, 1, and 0.1. The initial function $\phi^{(0)}$ is taken as the exact solution of the non-Coulombic term system (Eq. (3)) with $a_1 = a_2 = a_3 = \omega/2$, and the spin functions are taken as

$$\theta_1 = (\alpha_1\beta_2 - \beta_1\alpha_2)\alpha_3 \tag{11}$$

and

$$\theta_2 = \alpha_1(\beta_2\alpha_3 - \alpha_2\beta_3). \tag{12}$$

These two spin functions can represent any spin state of three-electron doublet state. The cfs generated are generally written as

$$\phi = A\left[r_1^{n_1} r_2^{n_2} r_3^{n_3} r_{12}^{m_{12}} r_{13}^{m_{13}} r_{23}^{m_{23}} \exp(-a_1 r_1^2)\exp(-a_2 r_2^2) z_3\ \theta\right] \tag{13}$$

where *m*s and *n*s are non-negative integers, and θ is either of Eq. (11) or (12). The sampling points $\{\mathbf{r}_\mu\}$ were generated in the Metropolis sampling method [15] so as to distribute along the density of the initial wave function. We used $N_s = 10^6$ sampling points.

In Table 1, the dependence of the energy on the number of the sapmling points are shown for ω=1 with the order n = 5 wave function. Even if the number of the sampling points increases, the energy does not change within 6-digits accuracy. This implies the wave function is almost independent of the sampling points.

Table 1. FC-LSE energy of the three-electron harmonium for ω = 1.0 calculated with *m* x 10^5 sampling points at order n = 5.

Number of Sampling points	FC-LSE energy (au)	σ^2 error
1 x 10^5	7.3397433	1.137 x 10^(-5)
2 x 10^5	7.3397343	1.622 x 10^(-5)
3 x 10^5	7.3397388	1.528 x 10^(-5)
4 x 10^5	7.3397392	1.371 x 10^(-5)
5 x 10^5	7.3397322	1.353 x 10^(-5)
6 x 10^5	7.3397353	1.343 x 10^(-5)
7 x 10^5	7.3397387	1.294 x 10^(-5)
8 x 10^5	7.3397395	1.271 x 10^(-5)
9 x 10^5	7.3397364	1.288 x 10^(-5)
10^6	7.3397336	1.277 x 10^(-5)

 ISBN: 978-84-608-6082-2

Table 2. The FC-LSE energies of the three-electron harmonium.

order n	M	ω=10	ω=1	ω=0.1
1	12	61.1362460	7.3433333	1.0758841
2	43	61.1373310	7.3397419	1.0609231
3	119	61.1381907	7.3397329	1.0596357
4	276	61.1384248	7.3397350	1.0594616
5	568	61.1385124	7.3397336	1.0594491
E(Cioslowski) [a]		61.1385255	7.3397411	1.0594492

[a] Ref. [3]

In Table 2, the calculated energies and the number of generated cfs are shown for different ωs. At order n = 1, the energy has only three-digit acuracy for ω=1, but as the order n increases the energies become accurate. At order n = 5, the energy for each ω agreed with Cioslowski's value to five digits [3]. For ω = 1 at order n =5, our energies are a little lower than Cioslowski's energy in the sixth digits . The energy at order n = 1 for ω=10 is accurate compared to that for ω=1. It is because when ω is very large the harmonium atom can be regarded as the non-Coulombic-term system, since the last term in Eq. (1) is negligibly small compared to the second term, and that the initial function (3) is already a good approximation for such system.

In Table 3, ten cfs that have the largest coefficients are shown, where the cfs have the form of (13) and the exp( ) part are eliminated for simplicity. The most important cfs are the initial functions themselves (Eq. (3)) , as expected, with the coefficients 0.336126. The third to 32nd cfs are all $r_{12}^{m_{12}} r_{13}^{m_{13}} r_{23}^{m_{23}}$ type functions without $r_1^{n_1} r_2^{n_2} r_3^{n_3}$ terms. This observation is similar to the two-electron harmoium case, where the exact wave function is written as a product of the exact wave

Table 3. Ten largest coefficients and complment functions for ω = 1 at order = 5 [a].

coefficient	spin function [b]	Multiplied terms
0.336126	1	
-0.336126	2	
-0.327530	2	r_{13}
-0.291100	2	r_{23}
0.291100	1	r_{13}
0.289951	1	r_{12}
0.283847	1	r_{12} r_{13}
-0.270102	2	r_{12} r_{23}
-0.248970	2	r_{13} r_{23}
-0.238449	2	r_{12}

[a] Each cf is normalized to unity, and the total wave function is normalized to unity. The cfs have the form of Eq. (13).
[b] see Eqs. (11) and (12).

 ISBN: 978-84-608-6082-2

fcuntion of the non-Coulombic-term system and the $(1+r_{12}/2)$ term, as shown in (4).

5. Summary

In this study, we applied the FC-LSE method to solves the Schrödinger equation of the two- and three-electron harmonium systems. The results are as accurate as Cioslowski's values to five digits. The most important cf is the initial function (3), *i.e.* the exact wave fcuntion of the non-Coulombic-term system, and the next important cfs are type of $\phi = A[r_{12}^{m_{12}} r_{13}^{m_{13}} r_{23}^{m_{23}} \phi_0]$ functions.

In the presentation, we will show some more details of the theory and the calculated results.

6. References

[1] N. R. Kestner and O. Sinanoğlu, Phys. Rev. **128**, 2687 (1962)
[2] W. Zhu and S. B. Trickey, Phys. Rev. A, **72**, 022501 (2005)
[3] J. Cioslowski, K. Strasburger, and E. Matito, JCP, **136**, 194112 (2012)
[4] M. Taut, Phys. Rev. A 48, 3561 (1993).
[5] M. Taut, K. Pernal, J. Cioslowski, and V. Staemmler, J.Chem. Phys. **118**, 4861 (2003)
[6] X. Lopez, J. M. Ugalde, L. Echevarria, and E. Ludena, Ohys. Rev. A. 74, 042504 (2006)
[7] S. M. Reimann and M. Manninen, Rev. Mod. Phys., **74**, 1283 (2002)
[8] P.-F. Loos, Phys. Rev. A, **81**, 032510 (2010)
[9] H. Nakatsuji, Phys. Rev. Lett. **93**, 030403 (2004).
[10] H. Nakatsuji, Phys. Rev. A, **72**, 062110 (2005).
[11] H. Nakatsuji, H. Nakashima, Y. Kurokawa, A. Ishikawa, Phys. Rev. Lett. **99**, 240402-1-4 (2007)
[12] H. Nakatsuji, Acc. Chem. Res., **45**, 1480-1490 (2012).
[13] Hiroshi Nakatsuji and Hiroyuki Nakashima, J. Chem. Phys., **142**, 084117 (2015)
[14] Hiroshi Nakatsuji, and Hiroyuki Nakashima, TSUBAME e-Science J., **11**, 8-12, 24-29 (2014).
[15] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953); B. L. Hammond, W. A. Lester, and Jr. Reynolds, Monte Carlo Methods in ab initio Quantum Chemistry (World Scientific, Singapore, 1994).

Acknowledgement

The authors thank Prof. J. Cioslowski for letting us know about his studies of harmonium systems. The author (H.N.) thanks Prof. E. Ludena for private scientific discussions about harmonium systems.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

ANALYTICAL STUDY OF LABOR MARKETS BASED ON GRAPH THEORY

M. Lloret-Climent[1], J. Nescolarde-Selva[1], H. Mora[2], M.T. Signes-Pont[2]

[1] *Department of Applied Mathematics, University of Alicante*
[2] *Department of Computer Science Technology and Computation, University of Alicante*

emails: miguel.lloret@ua.es, josue.selva@ua.es, hmora@dtic.ua.es, teresa@dtic.ua.es

Abstract

Graph theory is a fundamental tool in the study of economic issues, and the input-output tables are precisely one of the main examples of it. This mathematical approach allows to model economic systems and helps understanding on its overall functioning. The main contribution of this work is to provide a suitable model to makes possible applying mathematical tools, such as graph theory, to better understanding the way the labor market works. In this paper, it is performed an analytical study of the labor markets by means network analysis. This approach allows using the network concepts such as coverage, invariance, orbit and structural functions. It is defined supply-demand and competition functions as basic relationships functions on labor markets. Finally, an application example on labor markets is described and some conclusions are drawn about it.

Key words: mathematical methods, circular flow, networks models, applied mathematics, modelling

1. Introduction

The main idea behind the structural analysis is to identify the links or relationships that exist in input-output table [1]. It helps to find circularities in economic phenomena [2] and allows us to have a diversified and comprehensive vision of the analyzed economy.

 ISBN: 978-84-608-6082-2

The labor market [3] can be considered as an ecosystem in which networks or trade flows occur between income and work and where everything is interconnected [4]. Circular flow of income is a model used to explain, in a very simplified way, the basic functioning of the economy. This model represents an economic cycle of income and work in which are represented the resource and income flows between the entities involved in an economy [5]. Based on this theory, the circular flow of the economic activity is made through the market [6]. The key ideas behind this circular flow are the movement of factors and relationships among economic agents.

The search for a balance in the ecosystem of the labor market according to the above scheme is based on the balance between the needs of products and services, job offers and job demands. In this case, the companies always find professional profile needed for each job vacancy and every worker would work in the right job for him/her. However, this simplified model is not taking into account other aspects such as salary, working conditions or the workplace.

In this research we aim to interpret the labor market through networks theory that are represented by graphs and where characteristic concepts of the theory of chaos as cover, invariance and orbits interact with the circular flow concept.

2. Basic concepts summary

The concepts and relations introduced in this and following sections formalize a mathematical theory that aims to advance in the development of models to help understand economic phenomena of current society. The application of these mathematical concepts and relations to domain of the economy can create analogies that improve our understanding on the functioning of the market, on the interdependence relations of the entities and on the real flow of goods and services from one sector of the economy to another. From there, criteria will be created to build economic policy and to reduce the inaccuracies in calculating the national income.

Definition 1: Labor Market $S=(M,R)$ is the pair formed by object set M, determined by all people offer their skills to employers in exchange for wages, salaries and other forms of compensation. Participants in the labor market include any person x_i who is seeking to work for a wage and any person or company y_j that is looking for people to perform labor and a set of binary relations, so that $R \subset P(MxM)=P(M^2)$; that is, $\forall r \in R \ / \ r \subset MxM$ where $r=\{(x_1,y_1),(x_2,y_2),...,(x_i,y_i),.... /(x_i,y_i) \in \mathrm{MxM}\}$. ($P$ means "parts of a set").

 ISBN: 978-84-608-6082-2

Definition 2: In a labor market $S=(M,R)$, a set of companies $A \subseteq M$ $A = \{y_i\}_{i=1}^{n}$ is a circular flow if its elements satisfy the relationships $(y_1,y_2) \in r$, $(y_2,y_3) \in r$,..., $(y_{n-1},y_n) \in r$, $(y_n,y_1) \in r$, where $r \in R$. We assume that R is a set of relations between elements of M. For example, if r_c is the competition relation, $(y_i,y_j) \in r_c$ will signify that the companies y_i and y_j compete among themselves for more market share or if r_{sd} is the supply-demand relation, $(y_i,y_j) \in r_{sd}$ means that company y_i has outsourced company y_j performing some activity.

Definition 3: Let $S=(M,R)$ be a labor market and let the company $y_i \in M$. We will call the element in $P(M)$ formed of $f_c^n(y_i)\ \forall n \in N$ the orbit of y_i, and we will denote it as $Orb(y_i)$. In other words, it will include all the direct and indirect influences attained from y_i.

Definition 4: Let $S=(M,R)$ be a labor market and let the set of companies $A \in \mathrm{P(M)}$. We will call the element in $P(M)$ formed of $f_c^n(y_i) \quad \forall n \in N, \quad \forall y_i \in A$ the orbit of A, and we will denote it as $Orb(A)$. In other words, it will include all the direct and indirect influences attained from companies in A.

From these basic concepts, some relations among agents can be deducted within the theory framework to explain their behavior in the real word. The formulation, description and demonstration of these propositions are developed in the full manuscript. Deeper explanations about relation between the theory and how the labor market works will be also exposed in the full research. In addition, to help better understand the introduced concepts, an application example is detailed in a real case of the labor market.

3. Application Example & Discussion

An example illustrates a practical case of application of the theory proposed in a part of the labor market in a Spanish region. The analysis through the theory allows draw some conclusions on the labor market structure from the economic point of view.

In the example, the labor market will be segmented into three levels according the position of the firms in the circular flow of income. The results of the application of structural functions on agents and firms are described schematically in tables. The results and findings by applying the developed theory to the previous tables will be explained.

This example and a wide discussion about its results will be developed in the full research manuscript.

4. Conclusions

In this work a mathematical model for try to better explain the labor market mechanism has been proposed. The proposal described is based on previous works about graph and network theory. Some new concepts of the theory have been introduced. Orbits, coverage, invariant set and circular flow are defined between entities of the labor market.

As a result, the model predicts that decisions affecting the offers and demands for labor are not independent. Indirect relationships between offers and demands of job occur because the relations of the orbits of the companies. From these relationships, the structure of individual employment is determined by a causal chain that initially involves the supply market in a separate firm that eventually makes the offer. For example, the demand for a computer technician by a company depends on there being demand for these services from another company creating an indirect relationship to this ecosystem. Thus creating new labor opportunities and jobs is not depending exclusively on the contracting company but of the usefulness of their services to others firms.

Moreover, new relationships of competition according to the skills of workers are found. Specialized jobs are concentrated in specific firms and an increase of competition among unskilled workers occurs. These predictions of the theory are consistent with the evolution that is currently experiencing the labor market and new business strategies of companies.

5. References

[1] LEONTIEF. W. W. The structure of the American economy. *Sci Am*. **212**, (1965), 25-35.

[2] FONTELA E., AND GUZMAN J. Circulos viciosos y virtuosos del desarrollo económico. *Estudios de Economía Aplicada*. **21-2**, (2003), 221-242. (in Spanish).

[3] LANG, K. Racial Realism: A Review Essay on John Skrentny's After Civil Rights. *Journal of Economic Literature.* **53(2)**. (2015), 351-359.

[4] NORDHAUS, W.D., The ecology of markets, Proc. Natl. Acad. Sci. **89**, (1992), 843-850.

[5] MANKIW G., Principles of Macroeconomics, South-Western Cengage Learning, 6th ed. (2012), ISBN-10: 0324589999.

[6] LOPES, A. M., TENREIRO MACHADO, J. A. AND GALHANO, A. M. Multidimensional Scaling Visualization Using Parametric Entropy. *Int. J. Bifurcation Chaos,* **25**, (2015).

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Detached Eddy Simulation of Lateral Jet interaction flow

Jikui Ma, Yaofeng Liu and Yuwei Liu

China Academy of Aerospace Aerodynamics
emails: majikui2013@163.com, lyf545@sohu.com,
lyw642@126.com

Abstract

The numerical simualtion of the lateral jet interaction flowfield were carried out by detached eddy simulation. Two test models including the slot jet model and the ogive-cylinder body model are chosen to assess the performance of DES in jet intThis is a temperaction flow. Comparison was made with experimental data in terms of surface static pressure. Results showed that the DES model matched experimental data better than the traditional S-A model.

Key words: detached eddy simulation, jet interaction flow; separated flow

1. Introduction

The flowfield induced by a jet exhausting into a crossflow is a complex flow that has widespread applications in both military and civil industries, such as rocket motor thrust vector control systems, supersonic combustion, high speed flight vehicle reaction control system. A lateral jet issuing from surface body into a supersonic external flow acts as an obstacle to the flow, and creates a complex flowfield with shock/shock interaction, separated and reattached flow as well as complex spatial vortex structures.

As a consequence of these extensive applications, the lateral jet in supersonic crossflows has led to numerous studies up to now. Researchers have performed a vast number of computational studies for the problem of jet interaction. In the numerical analysis of complex lateral injection flowfields, one of significant issues is turbulence model in determining accuracy of the results.

Researches over the past decades have led to the development of numerous models which range from Reynolds-averaged Navier-Sotkes(RANS) to large-

 ISBN: 978-84-608-6082-2

eddy simulation(LES), to direct numerical simulation(DNS). The interaction flowfield involves complex 3D unsteady shocks, contact surfaces, turbulence and separation region, the traditional RANS failed to predic this unsteady characteristics and also perform poor in separated flow. LES and DNS attempts to resolve small scales of turbulence, the required computational resources increase dramatically, especially in high Reynolds number wall turbulence. Recently, detached eddy simulation(DES) which combines the strong points of RANS and LES approaches is received much attention. It gain huge success in massively separated flow.
In this paper, DES for calculation of lateral jet interaction flowfield was developed. The pressure distribution obtained from RANS and DES was also compared with the experimental results.

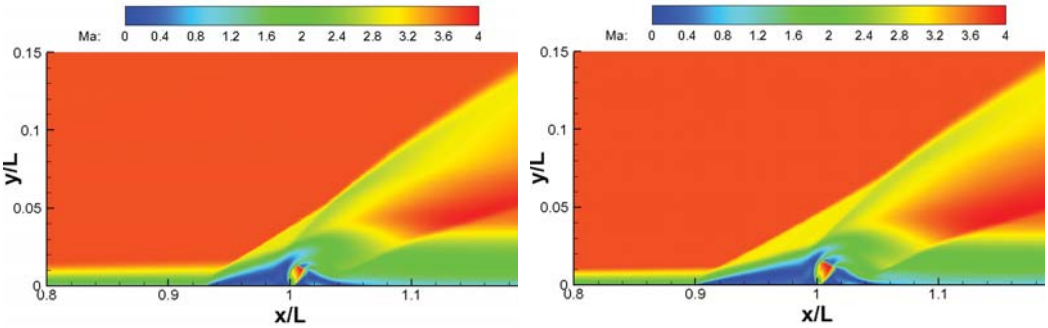
2. Numerical Methods

The computational results were obtained by solving three-dimensional compressible unsteady Navier-Stokes equations. The finite volume method was used to discretize the governing equation. Upwind-biased Roe flux-difference splitting scheme of second-order accuracy is implemented for the spatial discretization of the convective terms, the vanleer limiter is used to remove solution oscillations. Central difference scheme is used for the calculation of viscous terms, dual-time stepping method(inner time integration using the LU-SGS scheme) was employed for time discretization.
Detached-eddy simulation, as proposed by Spalart , is used in the paper. This is a hybrid model which combines the advantages of LES and RANS into one model. RANS is used in the boundary layer, where it performs adequately, and LES is used in the separated regions where its ability to predict turbulence length scales is important.

3. Results

The slot injection experiment of Aso et al. is numerical studied using DES as one of test models. The jet is vertically injected through a 1mm wide slot located 330mm behind the leading edge into a supersonic flow at Mach number 3.75. The exit Mach number of injector is 1.0 and the pressure ratio between mainstream and injection flow is 10.29.
Fig.1 shows the Mach number contours near the jet exit of S-A and DES respectively. Both methods can reproduce typical jet interaction flowfield characteristics， such as mach disk, separation shock, flow separation and recirculation. The separation length is shorter predicted by S-Athan the DES. We can also see from Fig.2 that S-Aunder-predict the pressure interaction region. The dissipation of the traditional RANS model is too large to predict the separated flow. The DES method improves this results obviously.

 ISBN: 978-84-608-6082-2



(a) S-A **(b) DES**

Fig.1 Mach number contours on the symmetric plane near the jet exit

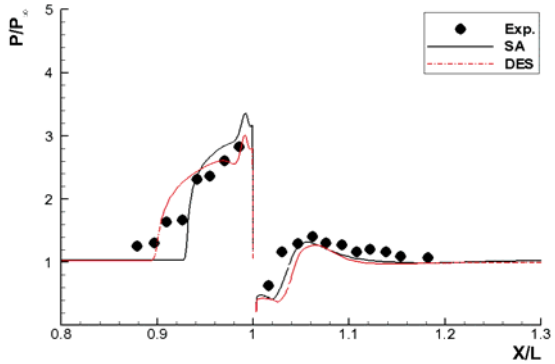


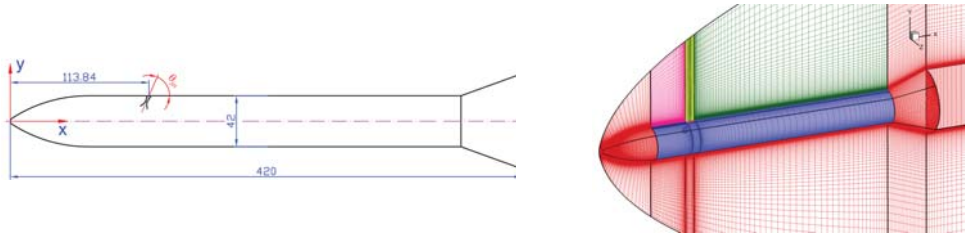
Fig.2 Comparison of surface pressure distribution

Another test model is an ogive-cylinder body with a jet nozzle of Mach number 3.0 as shown in Fig.3a. Figure 3b shows the grid structure for the model. The nozzle injection angle θ_{jet} is defined as the angle between the jet direction and the freestream direction. In this paper the injection angle θ_{jet} is fixed at $90°$. The freestream Mach number is 4.96 and total pressure is 2.39MPa. The pressure ratio P_j/P_∞ is 29.23. The prediction results are compared with the previous wind tunnel investigation. The DES based on S-Amodel is used.

Fig.4 shows the pressure contours on the symmetric plane and the pressure contours and streamlines on the surface near the jet. The DES method can reproduce the main interaction flow characteristics, such as bow shock ,separation shock , separation line and reattachment line.

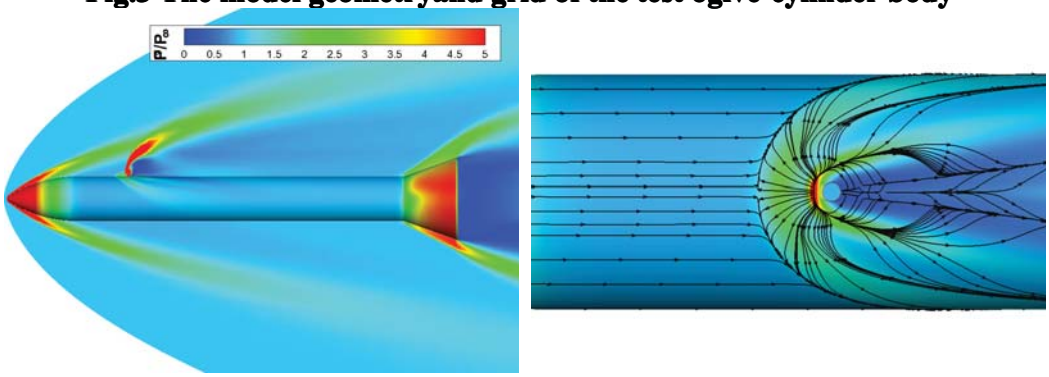
Fig.5 compares the pressure obtained by DES computation and experimental study along symmetric line . The DES result is agreed well with the experimental data. This method can predict the complex revolution body jet interaction flow.

 ISBN: 978-84-608-6082-2



(a) Geometry **(b) Grid structure of the model**

Fig.3 The model geometryand grid of the test ogive-cylinder body



(a) Pressure contours on the symmetric plane

(b) Pressure contours and streamlines on the surface near the jet

Fig.4 Flowfield obtained from DES simulation

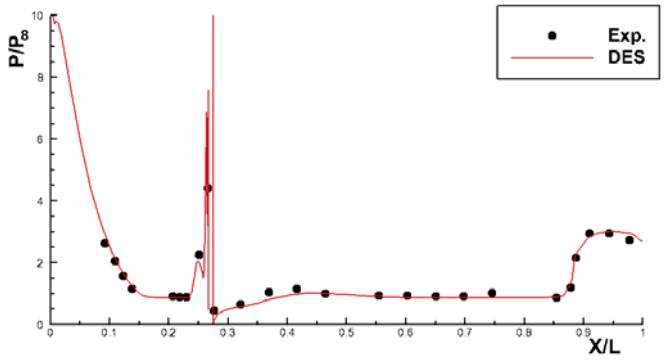


Fig.5 Comparison of surface pressure along symmetric line

3. Conclusions

Two typical interaction flow induced by the lateral jet exhausting into the high speed crossflow were numerical simulated with DES model. For the slot jet model DES model can correctly predict the mean flow structure and have an advantage over RANS model in separated flow. For the ogive-cylinder body the results also show that numerical prediction exhibits a fair agreement with the experiments and the main flow characteristics are reproduced well.

 ISBN: 978-84-608-6082-2

4. References

[1] E. E. ZUKOSIK, F. W. SPAID, *Secondary Injection of Gases into a Supersonic Flow*, AIAA Journal. **2** (1970) 1689-1696

[2] J. BRANDEIS AND J. GILL. *Experimental Investigation of Super and Hypersonic Jet Interaction on Configurations with Lifting Surfaces*, AIAA 97-3723.

[3] K. W. NAUMANN. *Stationary and Time-dependent Effects in the Near Interaction of Gaseous Jets and Supersonic Cross Flow*. AIAA 98-2972

[4] P. R. SPALART. *Annu. Rev. Fluid Mech.* 2009 181-202,

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Stress-strains contours in heterogeneous soils under arbitrary loads a the surface

Mena-Requena, M. R.[1], Morales, J. L.[2] and Alhama I.[1]

[1] *Civil Engineering Department, ETSIC*
Technical University of Cartagena
[2]*Department of Structures and Construction, ETSII*
Technical University of Cartagena

emails: mrmr0@alu.upct.es, joseluis.morales@upct.es,
ivan.alhama@upct.es

Abstract

Assuming linearly behaviour of a soil mass, for which elasticity process are ruled by two parameters, the modulus of elasticity and the Poisson´s ratio, numerical solutions of stress and strains distribution in 2-D, heterogeneous soils, subjected to loads of any form applied at any part of the surface, are studied. The models used are based on the network simulation method. Contours of the components of normal and shear stresses as well of deformations are presented for two typical applications where both qualitative and quantitative solutions reproduce the stress and deformation fields. Comparisons with the results derived from other techniques are made in order to test the reliability of the solutions.

Key words: elasticity, heterogeneous soils, stress and deformation contours, network method

1. Introduction

As known, the stress and strains distribution in a soil mass due to construction loads applied at the surface depends on the mechanical properties, geometry and uniformity of the soil, as well as the size and shape of the load. Under the assumption that the soil behaves as a homogeneous, isotropic, linearly elastic material the physical properties that influences the mechanical behaviour are the

 ISBN: 978-84-608-6082-2

modulus of elasticity, E, and the Poisson´s ratio, ν. Under these conditions, many of the solutions for special cases of the applied load obtained in the literature, as regards the stress distribution, have been analytically derived from Boussinesq and recovered in the work of Poulos and Davis [1]. The analytical solutions of these problems for heterogeneous soils, formed by a set of horizontal layers, or by arbitrary distribution of the loads at the surface are, however, more difficult to manipulate by engineers. They then treat to solve them by simulations in their own numerical software or by using standard commercial codes generally based on finite elements methods, such as MEFI [2].

In this work, numerical solutions for this kind of problems are obtained by network method, a tool extensively used in the last years in many problems of different engineering fields, such as microscopic tribology [3], chemical corrosion [4] and heat transfer [5] with successfully results. The network simulation method [6] is based on the equivalence between electric and physical models that result from spatial discretization of the partial differential governing equations that set the original mathematical model. Time, as in line methods, is retained as a continuous variable that is discretized by the simulation code used. Solutions of the network model are carried out in a standard free code of circuit simulation which provides, in fact, the exact solution of the model, so that the only errors in the method are due to the grid size chosen. The free code used in this work is Ngspice [7].

Firstly, the mathematical model (governing equations and boundary conditions) that follows the Navier´s formulation, in terms of displacements, for 2-D plane geometry, is presented. The design of the model may be found in Morales y col. [8]. Secondly, two applications are described: a two layer soil subjected to a triangular distribution load applied around the centre of the domain and a soil confined by a wall under a constant load at the corner. Contours of normal and shear stress as well as vertical and horizontal displacements are shown by interpolating the numerical solution. In addition, comparisons with the solutions of other numerical method are made in both application. Finally, conclusions are written in the last section.

2. Mathematical model

Governing equations for plane strain deformations are those of Navier. For a general domain shown in Figure 1, these equations in terms of displacements are

$$(\lambda + 2\mu)\frac{\partial^2 u_x}{\partial x^2} + \mu\frac{\partial^2 u_x}{\partial y^2} + (\lambda + \mu)\frac{\partial^2 u_y}{\partial x \partial y} = 0 \tag{2.1}$$

$$\mu\frac{\partial^2 u_y}{\partial x^2} + (\lambda + 2\mu)\frac{\partial^2 u_y}{\partial y^2} + (\lambda + \mu)\frac{\partial^2 u_x}{\partial x \partial y} = 0 \tag{2.2}$$

 ISBN: 978-84-608-6082-2

where λ and μ are the Lame´s coefficient and the transversal elasticity modulus, respectively. These properties are related to Young's modulus, E, and Poisson´s ratio, ν, by the expressions: λ=νE/[(1+ν)(1-2ν)] and μ=E/[2(1+ν)].

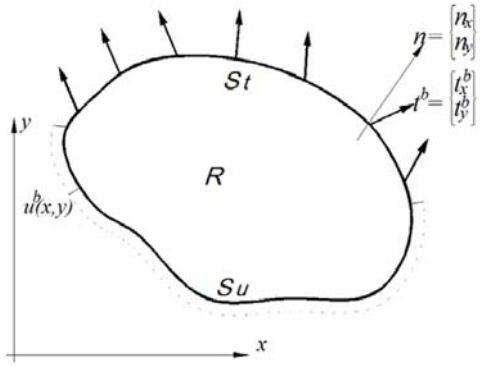


Figure 1. Domain of the problem for plane strain

Grouping the materials properties in terms of the new constants $C_1=(\lambda+2\mu)$, $C_2=\mu$ and $C_3=(\lambda+\mu)$, the above equations can be written in a more compact form, this is

$$(\lambda+\mu)\frac{\partial}{\partial x}\left(\frac{\partial u_x}{\partial x}+\frac{\partial u_y}{\partial y}\right)+\mu\left(\frac{\partial^2 u_x}{\partial x^2}+\frac{\partial^2 u_x}{\partial y^2}\right)=0 \tag{2.3}$$

$$(\lambda+\mu)\frac{\partial}{\partial y}\left(\frac{\partial u_x}{\partial x}+\frac{\partial u_y}{\partial y}\right)+\mu\left(\frac{\partial^2 u_y}{\partial x^2}+\frac{\partial^2 u_y}{\partial y^2}\right)=0 \tag{2.4}$$

As regards boundary conditions, the following equations applied according to displacement or stress specification at the surface

$$u_i = u_i^b \quad on \quad S_u \tag{2.5}$$

$$\sigma_{ij} n_j = t_i^b \quad on \quad S_t \tag{2.6}$$

$$t_x^b = \left[\lambda\left(\frac{\partial u_x}{\partial x}+\frac{\partial u_y}{\partial y}\right)+2\mu\frac{\partial u_x}{\partial x}\right]n_x+\mu\left(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x}\right)n_y \tag{2.7}$$

$$t_y^b = \mu\left(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x}\right)n_y+\left[\lambda\left(\frac{\partial u_x}{\partial x}+\frac{\partial u_y}{\partial y}\right)+2\mu\frac{\partial u_y}{\partial y}\right]n_y \tag{2.8}$$

3. Applications

3.1 Two infinite layers soil under triangular distribution load

Figure 2 shows the physical scheme of this problem. Two layers of different thickness and properties, with an enough length as to be considered horizontally infinite in the direction normal to the plane, are superposed with fix boundaries (lateral sides and bottom) except at the surface. The location of the load is symmetric while its distribution is triangular. Boundary conditions are pointed out in the figure.

 ISBN: 978-84-608-6082-2

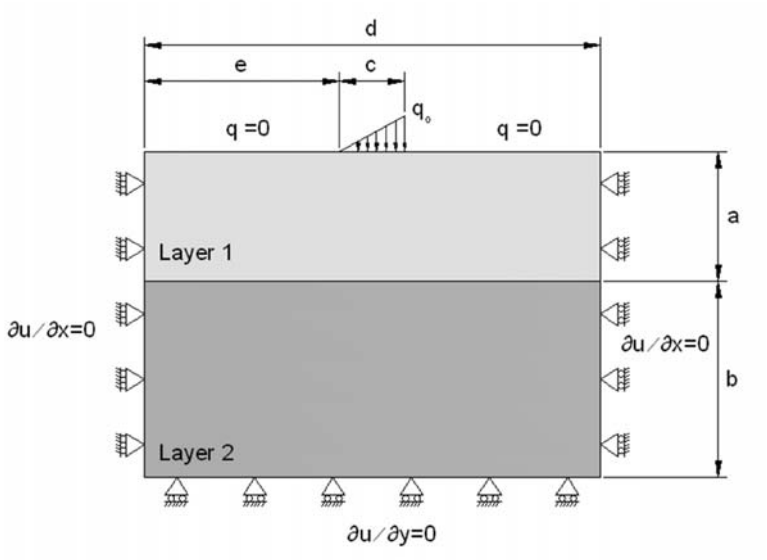


Figure 2. Physical scheme of the Application 1

The geometrical and physical properties of the problem are:

a = 20, b = 30, c = 10, d = 70 and e = 30 m; q_{max} = 4E3 kN/m^2
$E_{(layer\ 1)}$ = 5E5 kN/m^2, $\nu_{(layer\ 1)}$ = 0.4, $E_{(layer\ 2)}$ = 5E6 kN/m^2 and $\nu_{(layer\ 2)}$ = 0.2.

The grid size used is 70 (horizontal)×50 (vertical) cells or volume elements uniformly distributed (1m×1m each cell). Numerical solutions provide for the whole domain is interpolated to show graphs of x or y displacements at each point of the soil as well as graphs of continuous iso-lines of the components of the normal and shear stresses. The scale of displacements has been increased in the figure for a better appreciation.

Figure 3 shows the u_y and u_x displacement components of the whole domain in MATLAB format, while Figure 4 shows the stress isolines for the horizontal and vertical normal components as well as for the shear component, both in natural scale. It is interesting to note the partial symmetry of the displacement and stress, due to the triangular distribution of the load. Note that the higher stiffness of the layer 2 induces very small displacements in this layer and distorts highly the field of stresses at the boundary between layers. Three isolines of the horizontal components of the normal stress of zero value emerge in the solution. Two of them in layer 1, at both sides of the load, and one in layer two just down the load. However, the vertical components of the normal stress are continuous at the boundary between layers. Finally, the highest values of the shear stress emerge just down the corners of the loaded area.

 ISBN: 978-84-608-6082-2

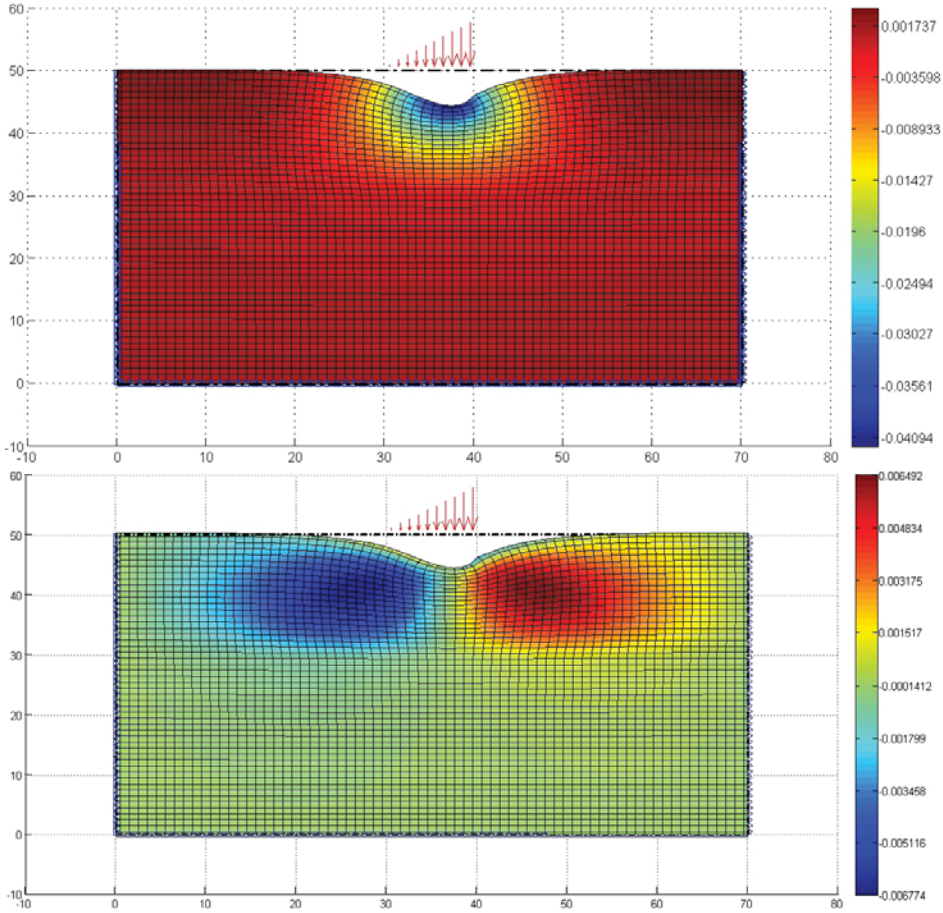
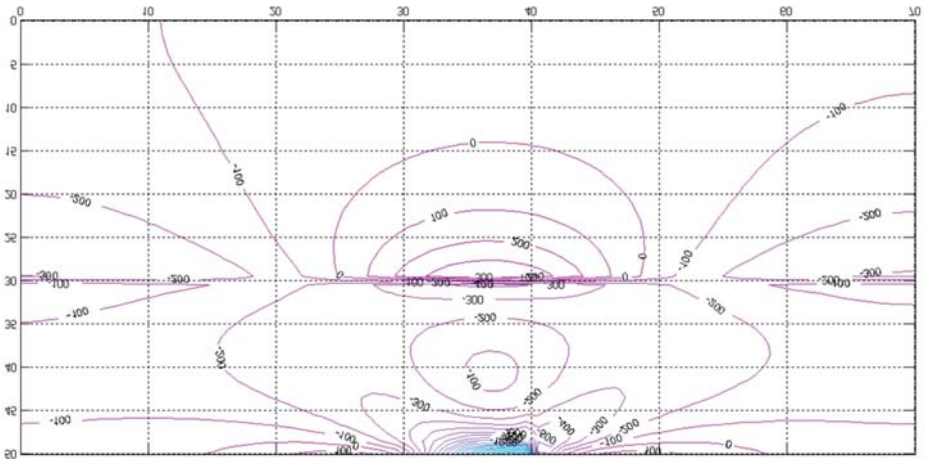


Figure 3. u_y and u_x displacement components, scale factor:137



 ISBN: 978-84-608-6082-2

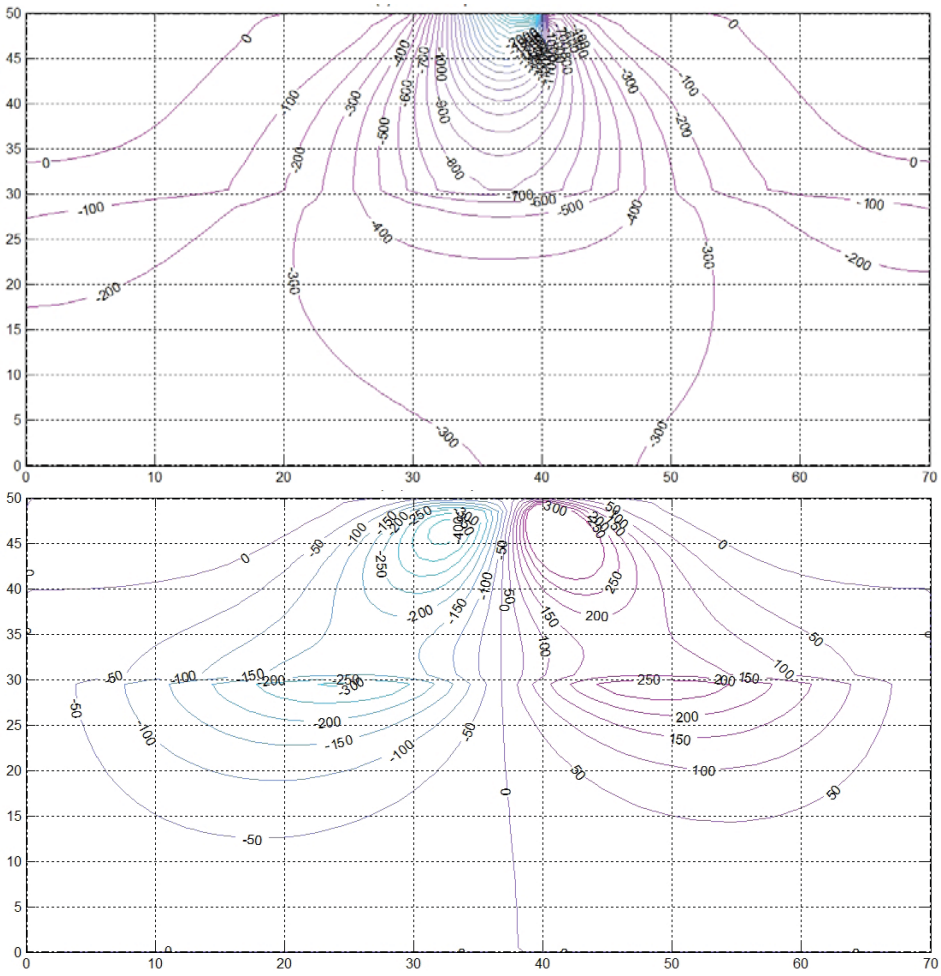


Figure 4. Stress isolines: σ_{xx} and σ_{yy} normal components (up and intermediate, respectively); σ_{xy} shear component (down)

3.2 Semi-infinite two layers soil under a constant load at the corner

Figure 5 shows the physical scheme of the problem with the boundary conditions. Geometrical and physical properties of the problem are:

i) Case 1, a = 10, b = 50, c = 10 and d = 110. q = 2E3 kN/m^2
ii) Case 2, a = 30, b = 30, c = 10 and d = 110. q = 2E3 kN/m^2

Young modulus and Poisson's ratio are those of the first application. The grid size for both cases is 55×30 (2m×2m for each regularly distributed cell at the domain). Numerical solutions for Case 1 show in Figures 6 and 7. The first depicts the u_y displacement, amplified 297 times, while the second shows the stress isolines for the normal and shear stress components. Similar comments to those made for the first application may be written. A large gradient of the horizontal component of

 ISBN: 978-84-608-6082-2

the normal stress emerges down the load, at the boundary between layers, due to the marked differences of the mechanical properties of these. In addition, shear stresses have also elevated at this boundary.

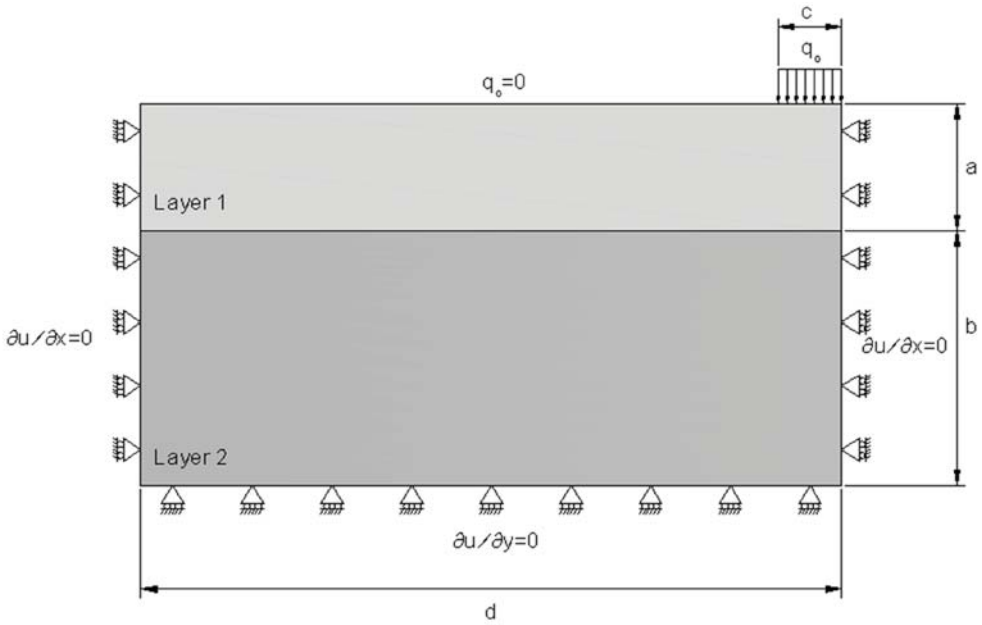


Figure 5. Physical scheme of the Application 2

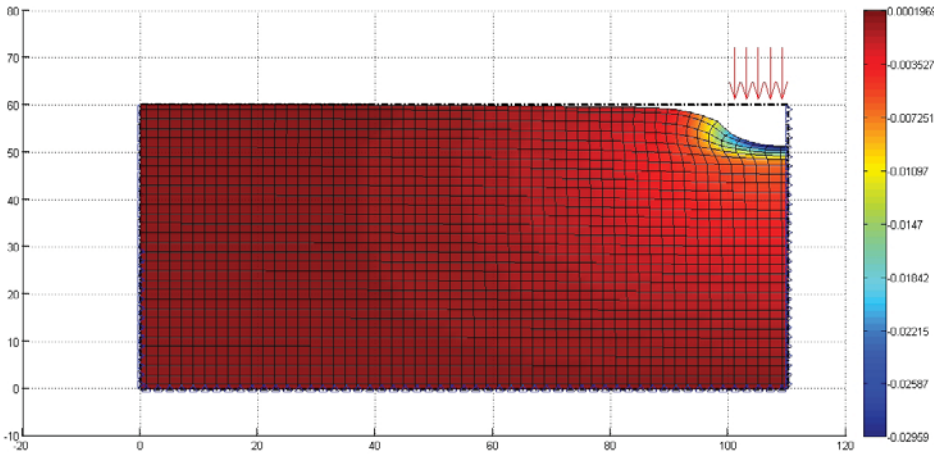


Figure 6. Case 1: u_y deformations, scale factor: 297

Numerical solutions of the y-displacement for Case 2 show in Figure 8. Note that the range of deviation in relation to the former case is larger because of the larger depth of layer 1.

 ISBN: 978-84-608-6082-2

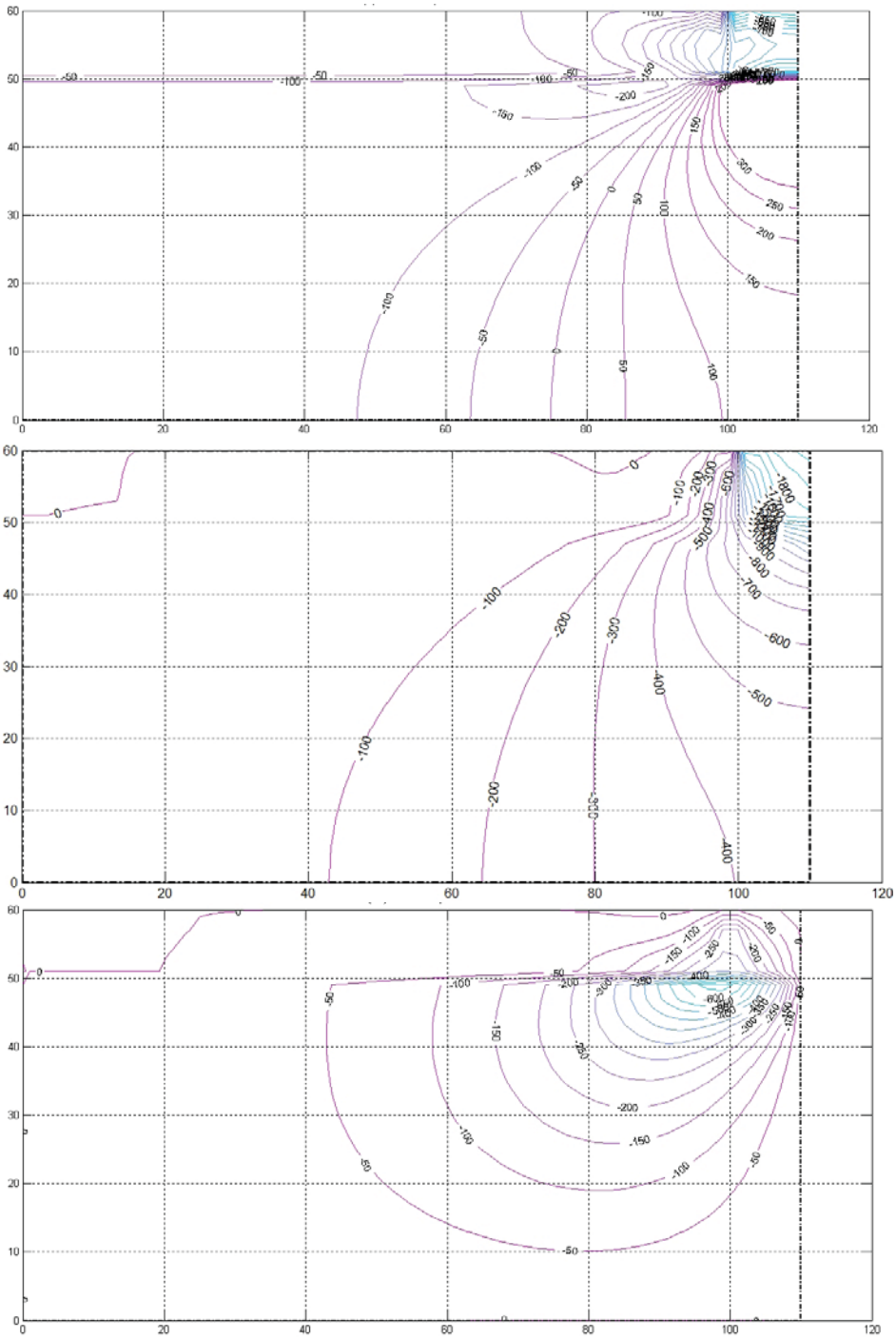


Figure 7. Case 2: Stress isolines: σ_{xx} and σ_{yy} components of the normal stress (up and intermediate, respectively). σ_{xy} shear component (down)

 ISBN: 978-84-608-6082-2

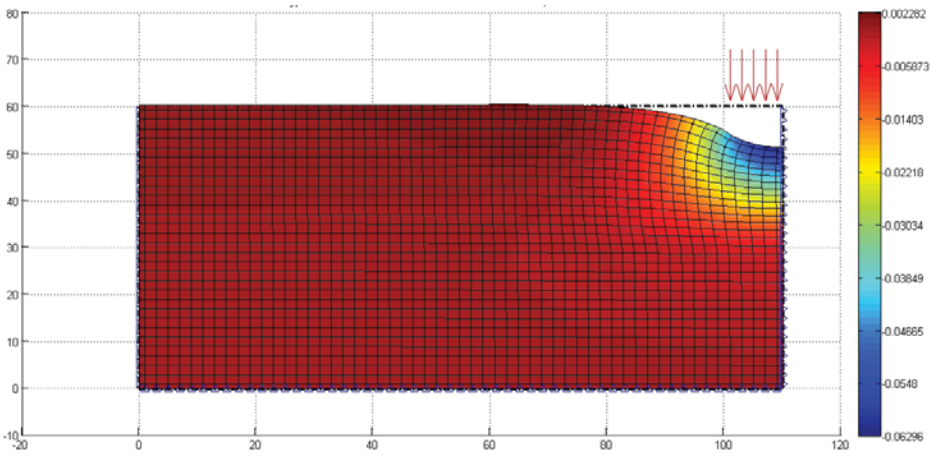


Figure 8. Case 2: u_y deformations, scale factor: 297

In order to check the reliability of these results, comparisons with the solution of MEFI for the y-displacements are carried out for the right point that defines the extension of the load. Table 1 shows the results that in any case do not overcome a relative error of 7 %.

	Application 1			Application 2			
				Case 1		Case 2	
	NSM	MEFI		NSM	MEFI	NSM	MEFI
x=e+c y=a+b	-0,0369	-0,0349	x=d y=a+b	-0,0296	-0,0316	-0,0630	-0,0659

Table 1. Comparison of maximum u_y-displacements between network method (NSM) and MEFI FEM code

4. Conclusions

The application of network method for the numerical solutions of the stress and displacement distributions in heterogeneous soil formed by two layers has proved to be an efficient and reliable tool. Standard mathematical routines allow the data be presented by isolines of constant stress or displacement values, which provides soil engineers with a fundamental information for the design and distribution of loads at the soil surface. No restrictions are required as regards the load distribution at the surface and the kind of heterogeneity. In all cases computing time, including data processing, is less than 1 minute for a PC i5.

References

[1] H. G. Poulos and E. H. Davis, *Elastic solutions for rock and soil mechanics*. Wiley, New York, 1974.
[2] MEFI, v1.1.2 (2015). The Finite Element Method in Engineering.
[3] F. Marín, F. Alhama y J.A. Moreno, *Modelling of nanoscale friction using network simulation method***,** CMC: Computers, Materials & Continua, **43** (2014) 1-19.
[4] J. F. Sánchez, Moreno Nicolás, J. A. and Alhama, F. *Numerical simulation of high temperature oxidation of lubricants using the network method.* Chemical Engineering Communications 202 (7) (2015) 982-991.
[5] E. Castro, M.T. Garcia Hernandez and A. Gallego, *Transversal waves in beams via the Network Simulation Method*, Journal of Sound and Vibration **283** (2005) 997-1013.
[6] C.F. González Fernández, *Network simulation method*, Ed. J. Horno, Research Signpost Trivandrum, 2002.
[7] Ngspice, (2013): mixed-level/mixed-signal circuit simulator.
[8] J.L. Morales, J.A. Moreno and F. Alhama, *Numerical solutions of 2-D linear elastostatic problems by network method*, CMES-Computer modeling in Engineering & Sciences **76**-1 (2011) 1-18.
[9] S. Timoshenko, J. N. Goodier, T*heory of elasticity*. McGraw-Hill, 1951.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Vibrational correlation formalism applied to internal conversion rate constants in metal clusters

Tzonka Mineva and Sandro G. Chiodo
[1] *Institut Charles Gerhardt Montpellier, CNRS/ENSCM/UM1/UM2, 8 rue de l'Ecole Normale,*
34296 Montpellier C´edex 5, France

emails:, tzonka.mineva@enscm, sandro.chiodo@enscm.fr

Abstract

In this paper we present a methodology to compute non-radiative internal conversion (IC) rate constants and its application to a series of small copper clusters (Cu_n, n=3, 6 and 9). The underlying theoretical background and details of the numerical implementation will be discussed. The need to quantify internal conversion rate constants possibly occurring in photoinduced catalysis are illustrated trough the analysis of the IC rate processes.

Key words: Internal conversion rates, TDDFT, hot-electron induced dissociation

1. Introduction

The interest to study non-radiative internal conversion is due to the recent renewed considerations of photophysical processes on noble nanoparticles in view of their potential applications for energy conversion and photocatalysis. For example, in the spherical noble nano-structures with a size > 2 nm, the localized surface plasmon, a collective oscillation of the conduction electrons, resonance occurs at the visible to near UV region of the spectrum. By an optical excitation in resonance of the surface plasmon band, the energy of electrons rises above the Fermi level to higher energy levels, creating so-called hot electrons. These hot electrons normally cool down very fast (in fs to ps time scale) due to scattering by phonons. If the hot electrons can be transferred to the anti-bonding orbitals of adsorbed molecules before they cool down, the dissociation reactions can be

 ISBN: 978-84-608-6082-2

enhanced by order of magnitudes [1]. It was shown that the small clusters have "molecular-like" optical behavior and the relaxation of the hot-electrons are slower (in the order of hundreds of ps) compared to the large particles.

2. Results and Conclusions

The method to derive the temperature dependent internal conversion rate constant (IC) is based on the path integrals of Gaussian type correlation function, originally proposed by Pollak et al [2] and further developed by Niu et al. [3]. The spontaneous emission rate can be evaluated, according to the Fermi golden rule, starting from the following general expression:

$$k_{ic}(T) = \frac{2\pi}{\hbar} \sum_{\upsilon_i \upsilon_f} P_{i\upsilon_i} \left| H'_{\upsilon_f \upsilon_i} \right|^2 \delta(E_{if} + E_{i\upsilon_i} - E_{f\upsilon_f})$$

where $P_{i\upsilon_i} = \prod_k \frac{e^{-\beta E_{i\upsilon_{ik}}}}{\sum_{\upsilon_k=0}^{\infty} e^{-\beta E_{i\upsilon_k}}} = \frac{e^{-\beta E_{i\upsilon_i}}}{Z_{i\upsilon}}, \quad E_{i\upsilon_{ik}} = (\upsilon_{ik} + 1/2)\hbar\omega_{ik}$

and $\delta(E_{if} + E_{i\upsilon_i} - E_{f\upsilon_f})$ is the Dirac-delta function, the indices *i* and *f* refer to the final and initial state, respectively. The ν_{ik} and ω_{ik} are the vibrational quantum numbers and frequencies of the first state, respectively. Using the Condon approximation, the matrix elements of the Born-Oppenheimer Hamiltonian can be expressed as:

$$H'_{\upsilon_f \upsilon_i} = \sum_l \left\langle \Phi_f \left| \hat{P}_{fl} \right| \Phi_i \right\rangle \left\langle \Theta_{f\upsilon_f} \left| \hat{P}_{fl} \right| \Theta_{i\upsilon_i} \right\rangle$$

In the present work the extension of this theoretical protocol, developed by Niu et al., was implemented and tested for transition metal clusters. The major computational difficulties to use the method proposed to study internal conversion processes in organic compounds is mainly because of the large number of d-multiplets in the transition metal clusters for which the excited state dynamics becomes computationally demanding. Despite computationally challenging, this formalism is a promoting-mode free rate containing Dushinsky rotation effect, so the mixing of the vibrational modes is explicitly considered and the method is fully analytical.

We applied this approach to quantify the internal conversion rate constants in minima structures of neutral copper clusters Cu_n (n=3,6,9) [4]. In the internal conversion processes the electronic excitation energy is transformed into the vibrational energy of the electronic ground state, which involves vibronic coupling, i.e., an electron-nuclear vibration interaction. If this interaction is large in electron-transporting materials, the dissipation of the energy should be large because the hopping electron couples strongly with intramolecular vibrations.

 ISBN: 978-84-608-6082-2

The IC values depend on the computational accuracy of the electronic excitation energies, ground, and excited states vibrational frequencies. It is therefore essential to provide also analysis about the performance of time-dependent Density Functional Theory that has been used to derive the excited state properties.

The most important conclusions can be summarized as follows: (i) computations of the non-radiative IC rate constants allow identifying the most probable relaxation pathways in the hot-electron induced relaxations, i.e. to identify the vibrational states to which the hot electrons might relax with the largest probability; (ii) the fast electron relaxation occurs when the H-H vibrational mode couples to the excitations involving mainly the transition between a Cu_3 molecular orbital with sp character and σ^* of hydrogen molecule; (iii) the results illustrate also the necessity to include the Dushinsky mixing effect in the IC description.

3. References

[1] S. MUKHERJEE, L. ZHOU, A. M. GOODMAN, N. LARGE, C. AYALA-OROZCO, Y. ZHANG, P. NORDLANDER, AND N. J. HALAS, *Hot-electron-Induced Dissociation of H2 on Gold Nanoparticles Supported on SiO_2*, J. Am. Chem. Soc. **136** (2014) 64-67.

[2] J. TATCHEN AND E. POLLAK, *Ab Initio Spectroscopy and Photoinduced Cooling of the trans-Stilbene Molecule*, J. Chem. Phys. **128** (2008) 164303.

[3] Y. NIU, Q. PENG, C. DENG, X. GAO, AND Z. SHUAI, *Theory of Excited State Decays and Optical Spectra: Application to Polyatomic Molecules*, J. Phys. Chem. A **114** (2010) 7817-7831.

[4] S. G. CHIODO AND T. MINEVA, *Application of Vibrational Correlation Formalism to Internal Conversion Rate: Case Study of Cu_n (n=3, 6 and 9) and H_2/Cu_3.* J. Chem. Phys. 142 (2015) 114311.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Combined use of Geogebra and 3D impression for Geometry learning

Lara Orcos[1] and Nuria Arís[1]

[1] *Department Education, Universidad Internacional de La Rioja (UNIR), Avda. Gran Vía 41, 26002, Logroño, (La Rioja) Spain,*

emails: lara.orcos@unir.net,nuria.aris@unir.net

Abstract

A high number of difficulties have been detected in the field of geometry related with students´ visualization skills at the last high school level. This study is an approach of an intervention in order to improve students' knowledge of geometry mathematical concepts throughout their association with the graphical representations in the second course of Science and Technology high school level. To carry out this purpose, several resources such as 3D manipulative resources and ICT resources (dynamic mathematics software) will be used to contribute to the development of students' spatial intelligence and a better understanding of mathematical concepts due to the fact that they are innovative methodologies that respond to the requirements of the students in the present-day society.

Key words: geometric visualization competence, graphical representation, dynamic mathematics software, manipulative resources, 3D impression.

MSC2000: 97D40, 97U99

1. Introduction

In relation to geometry, it has been found that students have problems when trying to understand geometry curriculum contents. Specially, the most typical problems occur in the study of vectors, planes and lines and their corresponding equations. This problem has been studied since decades ago, and studies such as Brihuega (1997) stated that the most important fact that induced the problems students had with geometry was that it was taught from an algebraic point of view instead of trying to develop the visualization competence.

The result of this way of teaching is that students just learn formulas and the lack of critical thinking is observed when the problem statement changes. The abstract reasoning does not take place in their minds and they do not look to the geometric basis that take place to the algebraic formula.

Some studies (Barrantes and Balletbo, 2012) consider that geometry is a mathematical area which evolves loads of benefits and it is of great interest in the field of education. It helps students to improve their visual abilities and develop their critical thinking. Moreover, geometry is of the interest of students due to the fact that it takes places of the real life, loads of geometrical figures can be found in the daily life. From a constructivism point of view, by which students built their own knowledge taking their motivation as the starting point, geometry is one that fields of knowledge more feasible to potentiate significant knowledge based on the previous concepts of the students.

It is important to mention that the origin of geometry was the aim of understanding the form of physical objects which are around us taking observation as an initial point of view. Trying to introduce this way of learning geometry in the classrooms can be the key to success.

Manipulative materials can be very useful in order to solve mathematical problems as they maximize students' creativity and abstraction capacity (Ledesma 2010).

Nowadays, thanks to ICT resources, such as games or special software the process of learning geometry becomes much more easy and motivating for the students and, what is better, it helps them to improve their spatial vision and so their abstraction ability.

In relation to 3D impression, it helps to illustrate the world throughout special models which are built for defined aims. It is important to take into account that motivation in human beings to things around us starts from our senses and the same happens with geometry. The motivation to geometrical figures starts in the eyes, so it seems required to let students use them in the classroom to learn geometry.

It is also important to emphasize that 3D impression can be a very effective resource to those students who have little visual ability and its use can help them to acquire the specific vocabulary in a significant way rather than in a memoristic one.

3D impression can be used by students of several courses from primary education to high level education and in several subjects and their economic costs are being reduced.

In order to cover a little more in relation to the research carried out on learning geometry it is important to remark some investigations which combine the use of the educational software GeoGebra with traditional methodologies (Iranzo and Fortuny 2009).

2. Objectives

The general objective of this study is to develop a geometry classroom intervention at the last level of Science and Technology high school education in order to help student to understand mathematical concepts throughout their relation to their graphical representations.

In order to get the main objective, some specific were stablished:
- To understand vectors, lines and planes equations through their graphical representation using both traditional methodologies and ICT resources.
- To perform students´ spatial ability that could help them with their following studies at university.
- To perform critical thinking that can help students to solve problems by their relation with their vital experiences.
- To maximize interdisciplinarity connecting geometry concepts in mathematics and drawing subjects.

3. Methodology

This study is based on the research about the integration of GoeGebra software and 3d impressions to solve analytic geometry problems.

The hypothesis to probe is whether the combined use of both methodologies, GeoGebra software and a manipulative technological one, 3D impressions, could enable students to understand the algebraic expression of vectors, lines and planes.

This intervention was carried out in the second course high level education in a group of 32 students of Science and Technology. To develop it, some activities has been proposed so as to help students to represent vectors, lines and planes equations using both GeoGebra software and 3D impression.

First of all a previous phase was carried in which it was important to study how the methodology was going to be carried out, the teacher availably and ICT performance, the resources, especially the 3D printer and its use. It was also required to test the abilities if students in using ICT resources.

During the planning phase it was important to carry out activities selection, their design and temporary implementation, as well as an evaluation methodology lined with the curriculum stablished.

Several sessions were required so as to approach to this planning phase. The methodology alternated 3D impression with GeoGebra. First of all 3D impression was carried out using the program Tinkercad to print son geometrical figures especially for the understanding of plane equations. Then it was carried out a brief explication of the software GeoGebra by using some easy examples. The learning process was carried out in groups to maximize collaborations between students with the aim that they could came up with the theoretical approaches.

4. Results and discussion

The most important result the study released was that all groups were able to come with the vector and line equations when working with these two methodologies.

There have been some important aspects that this study brought to light:
- Teacher are too much focused on solving algebraic problems rather than using graphical representations.
- Teaching mathematics in high school is so focused on the goal of getting good results to access to university instead of in developing important skills and competences such as spatial intelligence.
- Times and schedules sometimes enable the capability to implement these type of innovative methodologies into the classrooms.
- For students is so difficult to solve geometric problems due to the fact that they have little spatial capability so trying to abstract the concepts to solve the problems is even harder for them.
- Spatial intelligence is so important for the person because the world in which we move is in 3D, but the truth is that sometimes it is just released to the drawing subjects.

The work developed also brought to light several difficulties:
- The visualization process of the algebraic expressions of planes and lines is very hard for students.

- Students did not have problems in order to learn formula to calculate the distance between two lines or the angle between a line and a plane, but they do not use critical thinking with the results obtained because they had problems when trying to visualize them.
- The way by which teachers introduce geometry in the classrooms is focused in the algebraic point of view rather than in the visualization.

In scientific and technological fields, display ability can be very important and for this reason it is believed that the results of this study are really surprising due to the fact that students that opt form this high level studies al high school are supposed to study scientific or technological degrees in which spatial intelligence is very important.

The methodology proposed helps to improve the following aspects related to the mathematical competence.
- Critical thinking is improved so students can explain how they have come up with the solutions and why.
- The ability to visualize points, vectors, lines and planes in the space is highly increased improving the capacity to understand the required formulas to solve different problems more significantly.
- The ability to plan a problem has increase as they have potentiated their mental visualization.
- The combined use of the software GeoGebra and 3D impression as resources has been very helpful and highly motivating for students and have help them to improve other skills necessary in a scientific and technological world.

5. Conclusions

The combined use of manipulative resources, 3D impression, and a technological software has helped students to develop important skills such as spatial vision, digital competence and critical thinking in order to learn geometry concepts.

Moreover, the use of this methodology potentates the significant learning due to the fact that the students are the protagonists of their own learning process and that it enhances their motivation and creativity.

3D impression is an innovative resource not very use in the classrooms due to economic costs it requires, but taking into account that in the near future the acquisition of one of these resources will be more effective and the it can be used by students of different levels in several subjects it is proposed as a methodology that can introduce a load of benefits in the learning process.

Acknowledgement

This research was supported by Universidad Internacional de La Rioja (UNIR, http://www.unir.net), under the Plan Propio de Investigación, Desarrollo e Innovación 3 [2015--2017]. Research group: MOdelación Matemática Aplicada a la INgeniería(MOMAIN).

6. References

[1] M. BARRANTES LÓPEZ, AND I. BALLETBO FERNÁNDEZ., *Referentes principales sobre la enseñanza de la Geometría en Educación Secundaria*. Campo abierto: Revista de educación, **31**(2) (2012), 133-149.

[2] J. BRIHUEGA NIETO, *Las matemáticas en el bachillerato*. Suma, **25** (1997) 113-122.

[3] N. IRANZO N. AND J. M. FORTUNY, *La influencia conjunta del uso de GeoGebra y lápiz y papel en la adquisición de competencia del alumnado*. Enseñanza de las ciencias: revista de investigación y experiencias didácticas, **27** (3) (2009), 433-445.

[4] A. LEDESMA, *Aventuras y desventuras matemáticas de un folio DIN-A en el instituto*. Uno: revista de didáctica de las matemáticas, **53** (2010), 45-70

[5] A. RICHARD, *Textos clásicos y geometría dinámica: estudio de un aporte mutuo para el aprendizaje de la geometría*. Enseñanza de las ciencias: revista de investigación y experiencias didácticas, **28** (1) (2010), 95-112

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Design and dissemination of the MENTOR Tutorial Attention Plan in the School of Industrial Engineering of the Universidad de Valladolid

Ana M. Portillo[1], Marisa Fernando[1], Esperanza Alarcia[1], Laura Cuello[2], Pedro Díez[3], Sagrario Fernández[4], Nieves Fernández[4], José Mª García-Terán[5], Luis Carlos Herrero[3], Víctor A. Lafuente[6], Jesús Magdaleno[5], Mª Ángeles Martín[7], Fernando Martínez[3], José Manuel Mena[3], Cristina Pérez[3], Sara Pérez[4], Jesús Pisano[8], Virginia Reboto[9], Iván Rincón[4], Isabel Sánchez[9], Ana I. Tarrero[7]

[1]Departamento de Matemática Aplicada, Escuela de Ingenierías Industriales, Eii, [2]Departamento de Organización de Empresas y Comercialización e Investigación de Mercados, [3]Departamento de Tecnología Electrónica, Eii, [4]Departamento de Teoría de la Arquitectura y Proyectos Arquitectónicos, ETS de Arquitectura, [5] Departamento de Construcciones Arquitectónicas, Ingeniería del Terreno y Mecánica de los Medios Continuos y Teoría de Estructuras, Eii, [6]Departamento de Urbanismo y Representación de la Arquitectura, ETS de Arquitectura, [7]Departamento de Física Aplicada, Eii, [8]Departamento de Ingeniería Eléctrica, Eii, [9]Departamento de Química Analítica, Eii

emails: anapor@mat.uva.es, marisaf@mat.uva.es

Abstract

The European Higher Education Area has been a change in teaching methodology putting the student as the protagonist of their learning. This change modifies the role of teacher and student. The demand for more independent work by students, commits the university to create support and guidance systems. (Section 4.3 of Annex of RD 1393/2007 of October 29, which requires the existence in the degree courses of "accessible systems support and guidance of students once enrolled"). In this communication we present the work done to date in designing a Tutorial Attention Program, called "MENTOR

 ISBN: 978-84-608-6082-2

Program", aimed at all new students of the School of Industrial Engineering based on peer tutoring. This design includes: a comprehensive review of existing in other Spanish universities, study the roles and responsibilities of the members of the program (Tutor Teachers, Mentor Students, Tutored Students), setting the agenda for implementation in fixing the meetings of the various agents, enrolment periods, the selection of mechanisms for disseminating the program and materials design (logo, posters, ...) .

Key words: tutorial, transversal competences, social responsibility, permanent training

1. Introduction

Since the beginning of undergraduate degrees at the School of Industrial Engineering we have found that many new students have difficulties: complex schedules, three seats, student representation…

In this regard, 21 teachers from 9 different departments with teaching at the School of Industrial Engineering from the University of Valladolid (UVa), including 4 members of the management of the School of Industrial Engineering and General Secretary of the UVa, we are designing a Tutorial Attention Program called "MENTOR Program".

The program is based on peer support to facilitate speedy integration of new students into the university environment. We expect that upperclassmen, "Mentors", supervised by a teacher, "Tutor", orient and advise a group of new students, "Tutored" to help them in their academic and social integration in the university, and contribute the success of their studies.

The team of 21 teachers has extensive experience in teaching innovation, teaching in all degrees of the School of Industrial Engineering, different departments and areas of knowledge so it is an interdisciplinary project that focuses on coordination. It is geared primarily to the ongoing training of team members, the consolidation of our team, and social responsibility in our educational environment.

For each of the 10 groups of first course, two Mentors of the same degree that Tutored are sought. In total 20 Mentors are needed, although training courses would be extended to a somewhat larger number just in case there were unforeseen. With this project, Mentors will develop transversal competences in high demand by companies, such as leadership and oral communication.

The team of teachers has worked in sub-teams attending these steps:
Step 1: Find information on peer tutoring.

Step 2: Search training courses.
Step 3: Structuring and realize the MENTOR Program.
Step 4: Find ways of dissemination.

2. Objectives

The overall objective is design, for new students in the School of Industrial Engineering, a system of peer mentoring to facilitate their incorporation into university life, which will help them succeed in their studies.

The specific objectives are:

1. **Find information on different ways to perform actions tutoring**.
   They have been analyzed and studied tutorial action programs at the University of Burgos, of the Polytechnic University of Madrid, University of the Basque Country and the Carlos III University of Madrid. Seeing the dynamics of each of these programs we have determined our tutorial attention plan based on the characteristics of our school and our students.

2. **Search courses/workshops for "Mentors" so that they can develop their social skills, counseling and leadership.**
   We have raised three workshops for Mentors
   - On communication, teamwork and leadership. In our team there is a professor in the Department of Management and Marketing and Market Research who is an expert on these issues. She will give this course.
   - On the operation and organization of the University of Valladolid that will give the General Secretary, who is part of our team.
   - On the operation and organization of the School of Industrial Engineering who will teach a member of the Directorate (Director, Academic Secretary and two Assistant Directors are part of our team).

3. **Search and make courses/workshops for "Tutors" for enhancing their personal and professional training and a good follow-up mentoring process.**
   In January the team of teachers receives a course, "GUIDELINES FOR TUTORING among students" taught by Professors Almudena Ochoa and Piera Maresca, of the Polytechnic University of Madrid. With this course we decided how we were going to design our program, since the program Mentors UPM also serves a large number of new students and the characteristics of our School are very similar to those of ETSIDI University Polytechnic University of Madrid.

4. **Propose and make a good advertising system for the MENTOR Program.**
   A logo that identifies the MENTOR program (Figure 1), a website which has all the information and from which the registration of Mentors can be done, a poster (Figure 2) and a flyer (Figure 3) have been designed and diffused by the School. Mentor program has been spread too by Twitter and through Moodle.

   In addition, we have joined the Network of Mentoring University of Spain environments publicizing our project.

3. Dissemination Figures

Figure 1. Logo

Figure 2. Poster advertising

 ISBN: 978-84-608-6082-2

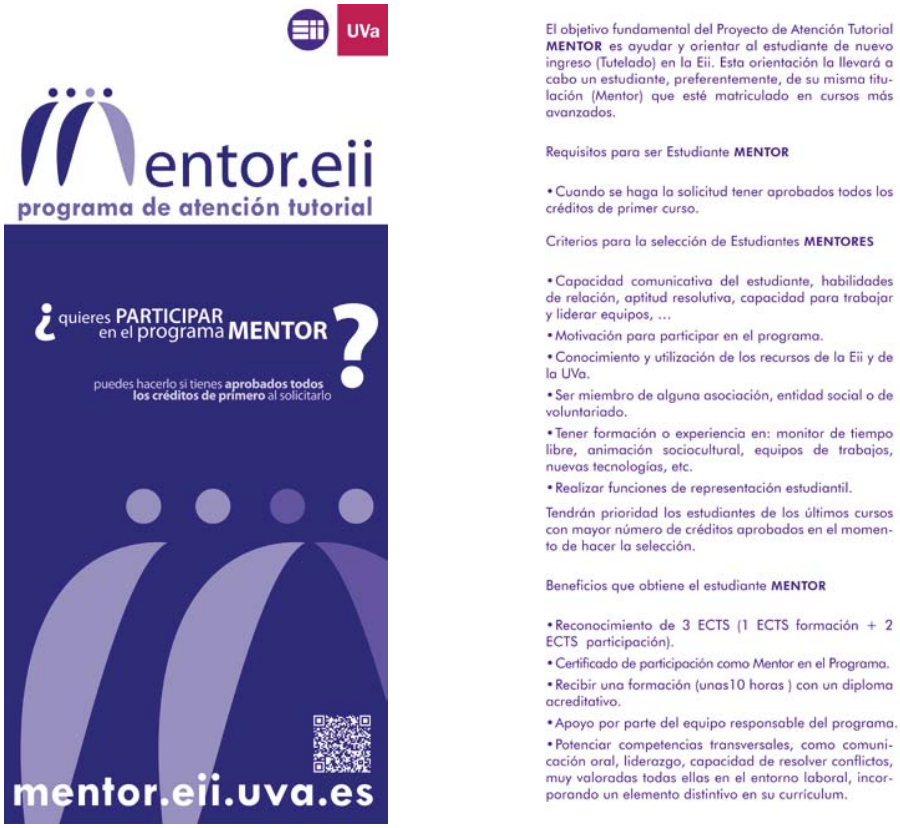


Figure 3. Flyer (both faces)

4. Schedule of the MENTOR Tutorial Attention Plan

April / May 2016

- Campaign to disseminate the new call of the Mentor Project.
- Mentors' registration in the website.

June / July 2016

- Meeting to pass the documents for the development of the Mentor 2016-2017.
- Mentors' selection. Assigning teams Tutor-Mentors-Tutored.

September to December 2016: Project Development

- First week of September:
  Seminar on general operation of the UVa (2 h).
  Seminar on general operation of the Eii (2 h).
  Seminar on leadership, teamwork and oral communication (6/8 h).

- Each month, two meetings Tutor-Mentors to prepare each meeting Mentors-Tutored.
- Each month, two meetings Mentors-Tutored. Mentors have to do a report of each meeting.
- In mid-December, Mentors will pass a satisfaction questionnaire about the Mentor Project to the Tutored.

March 2017

- Project closure meeting. Information about evaluation of it and suggestions for improvement.
- Satisfaction questionnaire about the Mentor Project, both Mentors and Tutors.

April 2017

- Delivery of the work end report from Mentors to their Tutors.

5. References

[1] Information concerning the UPM Mentor Project, http://www.etsidi.upm.es/ETSIDI/Estudiantes/AtencionAlAlumno/Proyecto+MENTOR. Last Access April 29, 2016.

[2] Information concerning the UBU Mentor Project, https://wwww.ubu.es/servicio-de-informacion-y-extension-universitaria/servicios-unidad-de-informacion/orientacion-y-tutoria-de-apoyo/programa-mentor. Last Access April 29, 2016.
Information concerning the Universidad Carlos III Mentor Project, http://portal.uc3m.es/portal/page/portal/cultura_y_deporte/orientacion/companeros. Last Access April 29, 2016.

[3] Information concerning the UPV Mentor Project, http://web.ua.es/es/ice/documentos/tutorial/material/ivjornada/pat-entre-iguales.pdf. Last Access April 29, 2016.

[4] A. OCHOA et al., *Claves de éxito para la implantación de un proceso de mentoría en una Escuela Técnica Superior de Ingeniería*, Actas de 23 CUIEET, 2015.

[5] C. SÁNCHEZ et al., *Proyecto Mentor en la Universidad Politécnica de Madrid: un sistema de mentoría para la acogida y orientación de alumnos de nuevo ingreso,* Sistemas, cibernética e informática, 6, n. 1, (2009), 64-71.

[6] M. FERNANDO et al., *Proyecto de Atención Tutorial MENTOR para la Escuela de Ingenierías Industriales,* VI Jornada de Innovación Docente de la Uva, 2016.

[7] A. PORTILLO et al., *Difusión para estudiantes Mentores, dentro del "Proyecto de Atención Tutorial MENTOR para la Escuela de Ingenierías Industriales",* VI Jornada de Innovación Docente de la Uva, 2016.

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Calculated Forecast for Technological Obsolescence in Computerised Tomography Equipment

Reyes-Santías, Francisco[1], Cadarso-Suárez, Carmen[2] and Espasandin, Jenifer[2]

[1] *Department of Business Management and Marketing, Universidad de Vigo*
[2] *Department of Statistics, Universidad de Santiago (USC)*

emails: francisco.reyes@uvigo.es, carmen.cadarso@usc.es, jenifer.esp.dom@gmail.com

Abstract

To estimate the useful life of Computerised Tomography Equipment (CT), with the aim of planning a budget the purchase of equipment and for the renewal of CT equipment within an already installed base so as to, therefore, maintain and ensure quality in providing an imaging diagnostic level in the National Health System.

The design of a prediction model provides an advance warning of the appearance of any technology leap that involves technological obsolescence in Computerised Tomography technology in use.

The starting data is made up by the different Computerised Tomography models commercialised since 1974 and that have provided a technology leap in CT technology.

It is necessary to know if the data distribution fits into a normal curve, because if this is not so the data will be transformed. A main components analysis in this methodology has allowed for a reduction in the number of variables on the survey-file in Computerised Tomography technology and facilitates subsequent work without a significant loss of information. The Log Binomial Regression Model has enabled probability calculations on answers (technology leap) to the different levels of stimuli (changes in variables, temporary development, detection system, imaging resolution and equipment power). Using a Discriminant analysis, the objective has been to estimate, based on time, the chances of a technological leap occurring.

Key words: CT, Medical technology, Obsolescence

1. Objectives set

To estimate the technological life of Computerised Tomography equipment (CT) with the aim of planning a purchasing budget to obtain equipment and renew the already installed base of technology equipment (CT) and therefore maintain and ensure quality in providing an imaging diagnostic level in the National Health System.

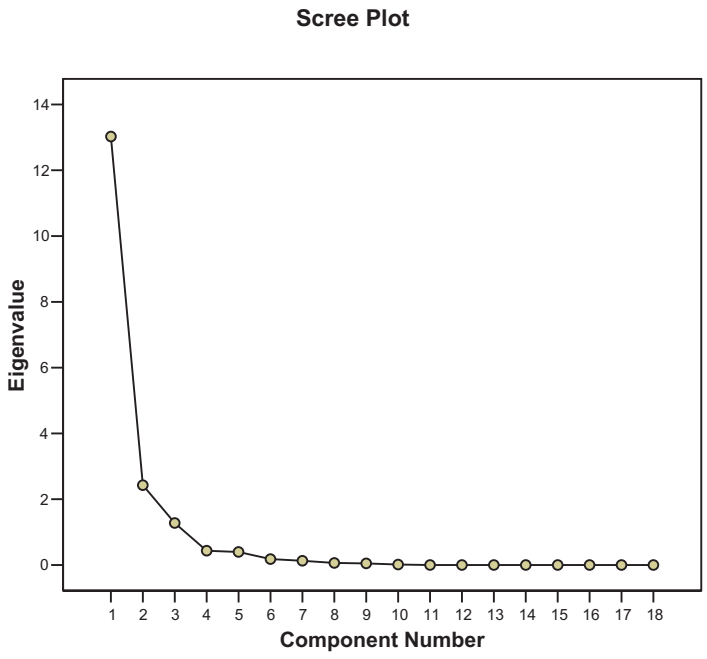
The design of a prediction model provides an advance warning of the appearance of any technology leap that involves technological obsolescence in Computerised Tomography Technology in use.

2. Results

Principal Components Analysis

The strategy followed has been to find a solution that enables us to explain the maximum percentage of variance and an acceptable parsimony of the model and to reduce the 18 descriptive CAT model variables; to set the explained percentage of variance in 92.9%, which is fulfilled in the first six factors in the present study and represented in graphic 1.

Graphic 1



 ISBN: 978-84-608-6082-2

A rotation of factors has been done and the varimax orthogonal rotation method has been used (axis turn orthogonally, in the same angle), which is intended to minimise the number of variables with high saturations in the same factor. The aim of this method is to increase the highest saturations in a factor, while decreasing the lowest for the factor to be easily interpretable.

The importance of each factor is evaluated considering the proportion of variance explained by the factor after rotation.

The 18 evaluated technical parameters in Computerised Tomography Technology have been grouped in three main components: *Detection System* (detector material, reconstruction matrices, reconstruction time, chest and abdomen scans, CT number range (HU), Processor, Maximum tube power (kW), scans ways) which explains 72.4% of the variance; *Imaging Resolution* (cutting thickness range, hard disk, selected kV, fine beam, thick beam) which explains 13.55% of the variance and *Equipment Power* (X-ray generator, number of data by turn or image, digital subtraction) which explains 7.1% of the variance.

Survival Analysis

The periods of follow-up in this type of analysis are almost always different, since the models of CT go incorporating to the study during all the period of observation, by what the last in doing it, shan been observed during a period of lower time that those that went in at the beginning. The time of failure of each model is measured from his date of entrance to the study. For each model of team has of a real time, that is the one who corresponds with the date in which this incorporates to the study until his last observation, and of a time "t" that it is the one who represents the time (in years, months, days, etc.) that was the model of team in follow-up.

It has employed the methodology of Kaplan-Meier, with which, the intervals determine by the occurrence of the event, that is to say, the probability of survival is calculated every time that occurs an event. The conditional probability of survival, that is the probability to be a survivor in the end of the interval conditioned to that it is a survivor in the beginning of the same for each model of team, calculates from the exact number of cases in risk when producing the event. It assumes that the instantaneous tax is zero during the interval between 2 events.

The methodology of Kaplan-Meier can be used to estimate the probability of survival above a period of time given. This method poporciona an estimator to be free of the event in the time t. For the calculations of the curve of survival, analysed a period of time of 34 years from 1974 to 2008. In this time entered 13

 ISBN: 978-84-608-6082-2

models of teams of CT. The times of survival (in years) of the 13 subjects were 3, 2, 1, 5, 3, 1, 2, 11, 4, 1, 5, 2.

The table 1 sample the table of life for these data, giving the proportion of survival in the times of survival no censored. The probability of survival remains in 0,8 until the time of the first event (3 years). The average of survival can obtain from the curve of Kaplan-Meier like the time in which the curve changes of a probability of greater survival of 0,5 to a minor of 0,5; in our investigation the average of survival of a model of CT are 4 years.

KAPLAN-MEIER

Sujeto	Status	Cumulative Survival	Standard Error
1	1	0,8	0,1789
2	0		
3	0		
4	1	0,4	0,2966
5	0		

Table 1

The curve of survival is drawn like a "staggered function": the proportion of survival remains constant between events. The times censored are indicated by marks on the curve of survival, which facilitate us the research of the times of survival of the subjects to which has not occurred them the event (fig. 2).

 ISBN: 978-84-608-6082-2

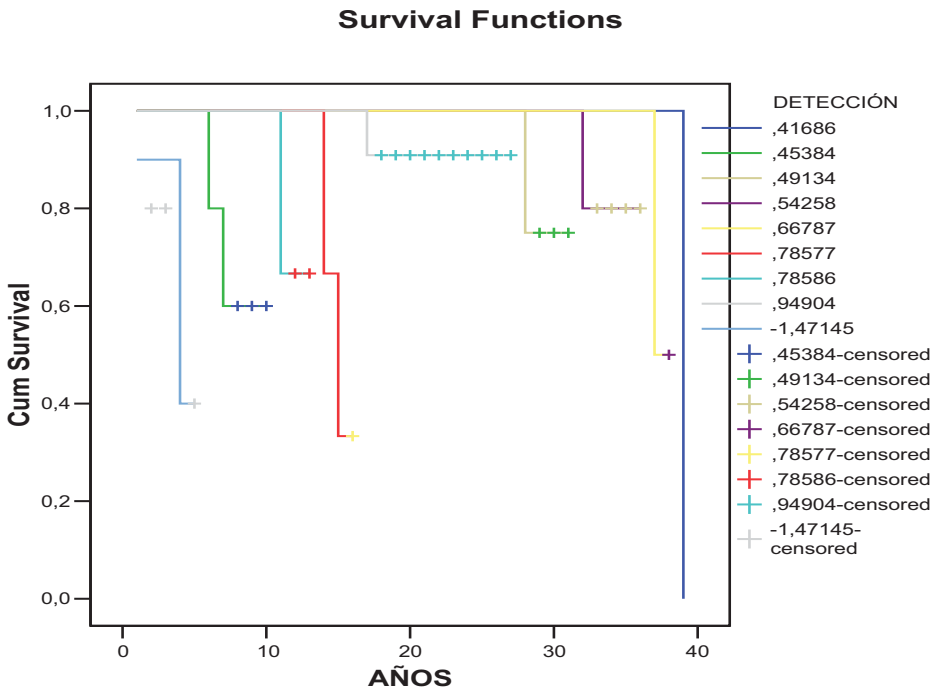


Figure 2

Analysis Logit

The characteristics shown by the design of a provision model in a technology leap demands the econometric specification is carried out through carefully chosen models, the endogenous variable to model is a categorical variable with various response alternatives. Within this modelling typology, Logit methodology adjusts to this aim using logistics as an adjusting function. The use of this function ensures that the estimation result can be interpreted as the occurrence probability of each endogenous variable alternative because the estimated values are always between the variation range 0-1.

Within the Logit modelling models of dichotomous answer and multiple answer models are distinguished, according to whether the endogenous variable to model has two or more answer alternatives with different specification in both cases. For a single case of dichotomous variable, in which there are only two possible answer alternatives (there is a technology leap or not), the endogenous variable is usually coded with 1 to represent the occurrence of the studied event or 0 to represent the non-occurrence.

The explanatory variables used in the model are the 3 main components previously calculated. The prediction model obtains a successful percentage of 76.9%. Table 2 presents the results of the Logit model for the current research,

 ISBN: 978-84-608-6082-2

showing its estimated coefficient β for the equation variables, typical error of β (S.T.), the Wald statistic, the degrees of freedom (df) and the p-value for the significance of the estimated coefficient (Sig.), the reason for the estimated advantages (Exp(β)) and the confidence interval for exp(β) at 95%.

Table 1.

VARIABLES	β	Std. Error	Wald	df	Sig.	Exp(B)
Intercept	-6,717	3,249	4,275	1	0,039	
DETECTION	-3,975	1,901	4,372	1	0,037	0,019
RESOLUTION	-3,766	1,814	4,312	1	0,038	0,023
POWER	-2,461	0,998	6,076	1	0,014	0,085
YEARS	0,431	0,194	4,904	1	0,027	1,538

The logistic regression allows us to approximate an assessment on the influence of each main component with the passing of time, the implementation of a technology leap, with its significant influence with positive signs of temporary evolution (0.430), and with a negative sign for the main component the detection system(-3.974), image resolution (-3.766) and equipment power (-2.460), as well as a technology leap expectative depending on the independent variables influence for temporary evolution (1.538) and main components, detection system (0.018) image resolution (0.023) and equipment power (0.085).

Discriminant Analysis

The monitoring periods of this type of analysis are almost always different given that the CT models incorporate into the study at different periods of the observation, in which case, the last in doing so have been observed during a lesser period than those that entered at the beginning. The failure time of each model is measured from the start date of each model's study. Every model has real time, this corresponds to its incorporation date until its last observation and a time "t" which represents time (in years, months, days and so on), the time in which each model was monitored.

The Discriminant methodology has been used so that the intervals are determined by the event occurrence, that is to say, the survival probability is calculated every time an event happens. The conditional probability of survival, this being the probability of being a survivor at the end of an interval which is conditioned by the fact that each model is already a survivor at the start of the interval, is calculated from the exact number of risk cases when the event occurs. This assumes that the instantaneous rate is zero during the interval between two events.

The Discriminant analysis can be used to estimate the probability of changing the CT model over a determined period of time and the factors influencing.

 ISBN: 978-84-608-6082-2

As in the Log Binomial, the explanatory variables used in the model are the 3 components calculated. The prediction model obtains a lower percentage of success than the Log Binomial, around 66.7%.

Table 2 presents the results of the Discriminant analysis, showing the estimated coefficient for the equation variables. The Canonical correlation shows a result of 0.481, meaning a good discrimination for the function. Moreover, the Lambda Wilkx shows a value of 0.769 and a p-value of 0.05.

Table 2.

Variables	Coefficients
DETECTION	1,822
RESOLUTION	2,250
POWER	1,147
YEARS	-2,888

The most important factor in influencing the change of technology seems to be the image resolution followed by the detection system and a negative sing for temporary evolution.

3. Relevance and practical use of results for the Health Area

The results of the present project will enable advance knowledge of the expectations of technological change in CT technology, allowing an advance in investment planning for this technology, for acquiring and installing the equipment in hospitals where this type of technology still does not exist and for the renewal of technology bases already installed.

The results of the present research can be validated by the application of other medical technologies, thus amplifying the impact of this research.

All this will have to be reflected by an investment budget plan for technology acquisition.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

QM/MM simulations of Au nanoclusters and glutathione ligands in water solvent

Víctor Rojas-Cervellera[1], Carme Rovira[1,2] and Jaakko Akola[3,4]

[1] *Departament de Química Orgànica and Institut de Química Teòrica i Computacional (IQTCUB), Universitat de Barcelona, Martí I Franquès 1, 08028 Barcelona, Spain.*
[2] *Institució Catalana de Recerca I Estudis Avançats (ICREA), Passeig Lluís Companys, 23, 08018 Barcelona, Spain.*
[3] *Department of Physics, Tampere University of Technology, P.O. Box 692, FI-33101 Tampere, Finland.*
[4] *COMP Centre of Excellence, Department of Applied Physics, Aalto University, FI-00076 Aalto, Finland.*

email: jaakko.akola@tut.fi

Abstract

Quantum mechanics / molecular mechanics (QM/MM) simulations have been performed to study the effects of aqueous solvent and biological ligands onto structural and electronic properties of thiolate-protected $Au_{25}(SR)_{18}^-$ clusters. The nanocluster structure experiences modest changes in the solvent, which are seen as flexibility ("fluxionality") in the Au core. The glutathione ligands shield the metallic core while distorting its symmetry via sterical hindrance effects. Our results demonstrate that the previously reported agreement between the calculated HOMO-LUMO gap of $Au_{25}(SR)_{18}^-$ and the optical measurement is due to cancellation of errors: The underestimation of the theoretical band gap has been compensated by missing solvent. For the solvated nanocluster with glutathione ligands, a hybrid DFT functional results in a HOMO-LUMO gap value of 1.5 eV, in good agreement with optical measurements. These results show that the effect of ligands/solvent should be included for a proper comparison between theory and experiment.

Key words: gold clusters, glutathione ligands, density functional theory, molecular dynamics, QM/MM

 ISBN: 978-84-608-6082-2

1. Introduction

Gold nanoclusters (AuNC) with protecting thiolate ligands have been extensively used as carriers of biological molecules, such as DNA, antibodies and specific proteins [1-3]. Various AuNCs of different sizes and a variety of protecting ligands for the gold core have been synthesized recently [4]. Among them, $Au_{25}(SR)_{18}^-$ is one of the smallest ligand-protected gold cluster for which the X-ray structure is known [5]. It is spherical and symmetric, constituted by an icosahedral Au_{13} gold core and six $Au_2(SR)_3$ dimeric staple motifs protect the gold core in an octahedral arrangement (Figure 1) [6]. Recently, structures of even smaller ligand-protected gold clusters have been predicted computationally [7].

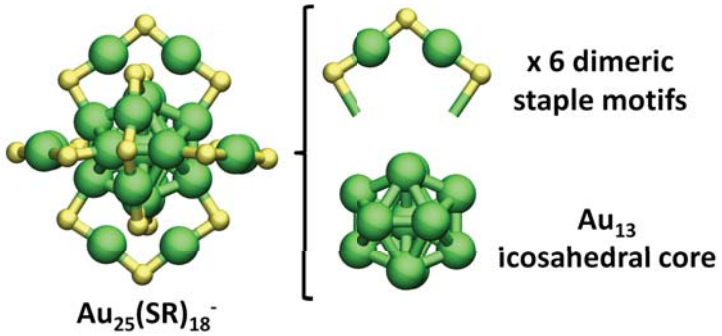


Figure 1. *Description of the $Au_{25}(SR)_{18}^-$ structure. The organic ligands attached to the sulphur atoms are not shown for clarity. Six dimeric staple motifs (SR-Au-SR-Au-SR) at the Au-S interface are the protecting units that are bonded to the Au_{13} icosahedral core. Gold and sulphur are shown green and yellow, respectively.*

Different hydrophobic alkanethiolates and chiral ligands have been used as ligand units [8-10]. Therefore, there may be nonspecific hydrophobic interactions that interfere in the binding target if these AuNCs are used to bind biological molecules. Some functionality needs to be introduced to the nanocluster to eliminate such interactions and to selectively bind the target molecules. Glutathione (GSH) is a tripeptide formed by a glutamate, a cysteine and a glycine (see Figure 2 for atomic structure). It is often used as a ligand because it is able to selectively bind proteins such as glutathione-S-transferase [3] and single chain antibody fragments [2]. The GSH amino acid sequence can be written as γ-Glu-Cys-Gly, and it has 2 asymmetric carbon atoms which introduce chirality. At physiological pH, both carboxylic acids are deprotonated, *i.e.* negatively charged, whereas the amino group is protonated, *i.e.* positively charged. Therefore, the total charge of a GSH ligand is -1.

It has been shown that $Au_{25}(GSH)_{18}^-$ is very stable [11], but its crystallographic structure has not been resolved. GSH is a bulky biological ligand comprising

 ISBN: 978-84-608-6082-2

several charged groups and its presence is expected to modify the structural and electronic properties of AuNC with respect to small alkanethiols. We present here QM/MM simulations of the structure and electronic properties of two clusters of composition $Au_{25}(SR)_{18}^-$ (R = glutathione or alkanethiolate) in vacuum and aqueous solution [12]. For simplicity, we have used methanethiolate (*i.e.* the smallest alkanethiolate) as a model of alkanethiol ligands, but this system already captures the essential chemistry at the Au-S interface layer. The simulated systems are presented in Figure 2, and the division between QM and MM domains is highlighted.

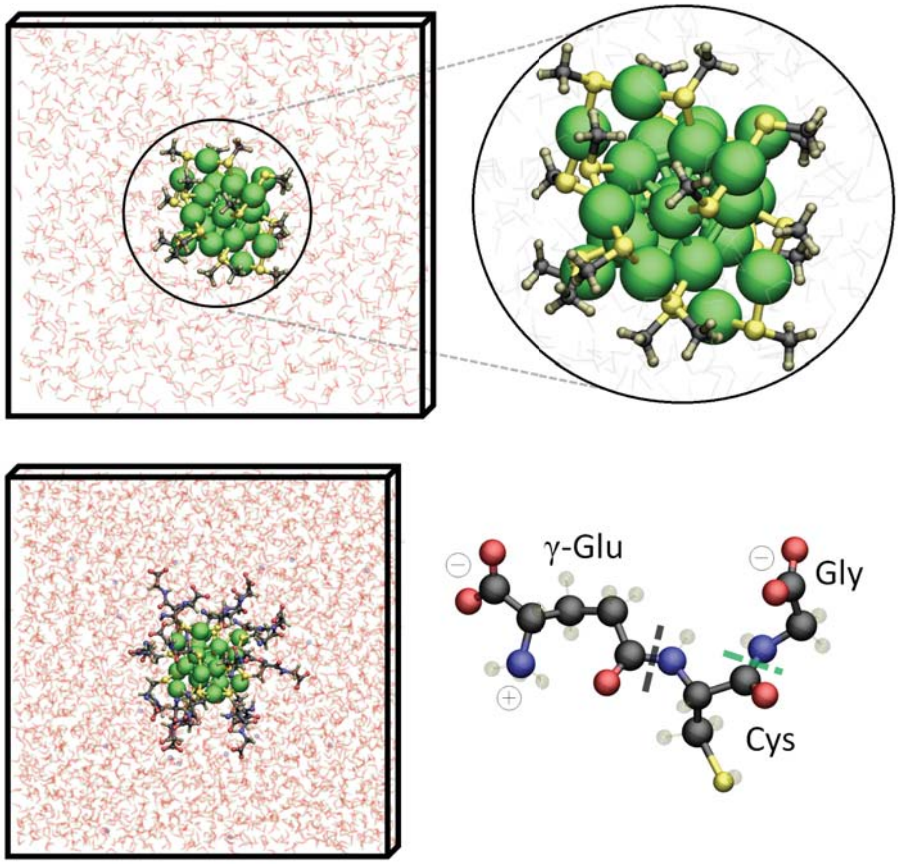


Figure 2. *The solvated $Au_{25}(SCH_3)_{18}^-$ (top) and $Au_{25}(GSH)_{18}^-$ (bottom). For $Au_{25}(SCH_3)_{18}^-$ the QM region is zoomed on the right, where the QM zone is shown in color (Au in green, S in yellow, C in black and H in grey). For $Au_{25}(GSH)_{18}^-$, the ball & stick representation of a GSH peptide (negatively charged) is shown on the right. Hydrogens are shown transparent for clarity. The black dashed line shows the γ-peptide linkage between Glu and Cys. The green dashed line shows the normal peptide linkage between Cys and Gln.*

 ISBN: 978-84-608-6082-2

2. Methods

Prior to the QM/MM molecular dynamics, MM simulations with classical potentials were performed, up to 8 ns, using the NAMD software [13]. Ligands were modelled with the FF99SB force field [14], and water molecules were described with the TIP3P force field [15]. Snapshots of the MM equilibrated systems were taken for the subsequent QM/MM simulations. The QM/MM interface is modelled by the use of monovalent carbon pseudopotentials which saturate the valence of the QM atoms at the border (linking atoms) [16]. The electrostatic interactions in the interface of the two regions are described in reference [17].

For the QM part, we perform *ab initio* molecular dynamics (AIMD) simulations based on the density functional theory (DFT) of electronic structure. AIMD simulations were performed within the Car-Parrinello approach [18] and using the CPMD package [19]. A fictitious electron mass of 600 a.u. and an integration time step of 0.12 fs guaranteed a good control of the adiabaticity for ionic and electronic equations of motion in all the systems where QM or QM/MM molecular dynamics simulations were performed. The simulation temperature was set to 300 K by coupling it to a Nosé-Hoover chain thermostat for the ionic degrees of freedom [20]. The Kohn-Sham orbitals were expanded in a plane wave (PW) basis set with a kinetic energy cutoff of 90 Ry. The exchange-correlation functional employed the parameterization by Perdew-Burke-Ernzerhof (PBE) [21]. Furthermore, a hybrid functional (PBE0 [22]) was used to compute HOMO-LUMO gaps for the optimized structures in order to evaluate the effect of including a portion of the exact exchange interaction on the electronic properties. The electron-ion interactions were described by using pseudopotentials which take into account the combined effect of nuclei and core electrons onto valence electrons.

Three model systems were prepared for the simulations: (1) $Au_{25}(SCH_3)_{18}^-$ in the gas phase; (2) $Au_{25}(SCH_3)_{18}^-$ in an aqueous solution; and (3) $Au_{25}(GSH)_{18}^-$ in an aqueous solution. The structure of the $Au_{25}(SCH_3)_{18}^-$ system was optimized with DFT, followed by room temperature AIMD simulations for 7.5 ps using the CPMD program. An explicit water solvent was considered for the last two systems, which were modelled with the CPMD QM/MM approach for 7.5 ps, each.

3. Results

The solvent does not affect the global structure of the $Au_{25}(SCH_3)_{18}^-$ framework significantly, as the Au_{13} core and the six (protecting) staple motifs are preserved. However, the cluster expands slightly: the largest difference between the isolated

 ISBN: 978-84-608-6082-2

and the solvated $Au_{25}(SCH_3)_{18}^-$ at 0 K appears for the Au(a')-Au(c) distance (0.11 Å or 3.2 %) which involves the outer core gold atoms (icosahedron vertices, Au(a′)) and those of the staple motifs (Au(c)). During the MD simulation (300 K), the average values of the Au-Au bonds show a further increase up to 0.10 Å (for Au(a')-Au(a')), while the S-Au bonds remain almost intact (within ± 0.03 Å).

Figure 3 displays the distribution of interatomic distances for the solvated $Au_{25}(GSH)_{18}^-$ at room temperature (QM/MM simulation). For the Au-Au bonds (center (a), 12 core vertices (a′) and 12 staple (c) atoms), the distributions are very broad, which demonstrates the flexibility of these bonds, termed previously as "fluxionality" for bare Au clusters [23]. Hence, the introduction of bulkier ligands can easily induce changes in the Au_{13} core. The Au-S bond distances show two distributions depending on the type of the Au atom (core or staple) with markedly different locations and shapes.

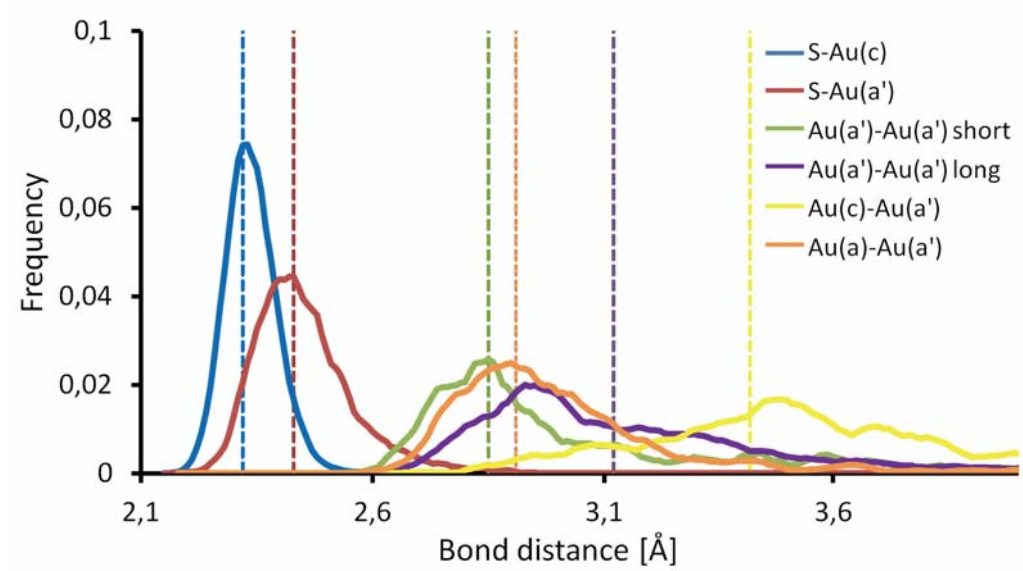


Figure 3. *Histogram of the six main distances that define the structure of $Au_{25}(GSH)_{18}^-$ during the 7.5 ps QM/MM simulation at 300 K. The dashed lines show the optimized values for each distance. Note that the Au(c)-Au(a') bonds are weak ("aurophilic"). The calculated radii of gyration are 4.12 and 4.10 Å for the full Au_{25} cluster or 1.49 and 1.44 Å for only the Au_{13} gold core, for ligands being GSH and SCH_3, respectively.*

Comparison between the AuNC structures with the two ligands (thiolate and glutathione) shows that the overall structure of AuNC is not significantly distorted. However, marked variations are found at 0 K for the Au(a')-Au(a') bond distances (GSH), which increase up to 0.08 Å in comparison with the methanethiolate ligands. Furthermore, the increase of the Au-Au bond distances is even more pronounced at 300 K in comparison with $Au_{25}(SCH_3)_{18}^-$ in vacuum (0.09-0.14 Å). This is due to the solvent effects and steric hindrance of the bulky

 ISBN: 978-84-608-6082-2

ligands, which modify the Au-Au bonding via mechanochemical coupling. For GSH ligands, the Au_{13} core expands slightly to accommodate better the bulky ligands, and the first shell water solvent causes electrostatic effects on the electron density. The Au(a')-S-Au(c) angle decreases due to the formation of hydrogen bond interactions among the GSH ligands.

Table 1. *Hirshfeld charges of the five atomic types for Au and S. The labelling of sulphur atoms distinguishes between the staple ends (S) and apex atoms (S_{ap} in the middle).*

System	a	a'	c	S	S_{ap}
$Au_{25}(SCH_3)_{18}^-$ in vacuum	-0.673	-0.176	0.455	-0.263	-0.423
Solvated $Au_{25}(SCH_3)_{18}^-$	-0.649	-0.182	0.445	-0.307	-0.436
Solvated $Au_{25}(GSH)_{18}^-$	-0.591	-0.176	0.400	-0.294	-0.481

Atomic (Hirshfeld) charges were computed from the valence electron density to analyze the effect of the solvent and the ligand type on the electronic structure of the AuNCs. Table 1 displays the computed values for the different types of gold and sulfur atoms and Figure 4A displays their changes using a radar chart.

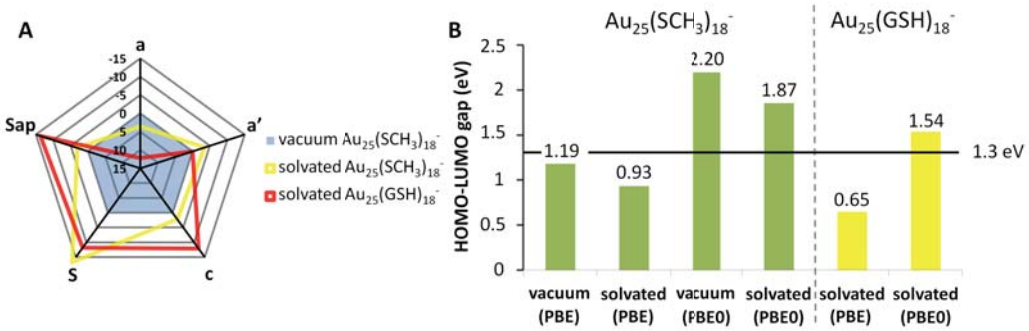


Figure 4. *(A) Change of the atomic Hirshfeld charges for the most representative atoms of $Au_{25}(SCH_3)_{18}^-$ and $Au_{25}(GSH)_{18}^-$ solvated in water with respect to the ones obtained for the isolated $Au_{25}(SCH_3)_{18}^-$ (blue area). (B) HOMO-LUMO gap of these systems with PBE and PBE0 exchange-correlation functionals. Green bars correspond to $Au_{25}(SCH_3)_{18}^-$, whereas yellow bars refer to $Au_{25}(GSH)_{18}^-$.*

Introducing the solvent makes sulfur atoms more negative, and the corresponding charge is partially transferred from the negatively charged central Au atom (a) which has an effective charge of -0.65e. The charge transfer effect gets more pronounced for S_{ap} (apex, -0.48e) when the ligand is GSH, and hence the staple motifs become less electrophilic. The Au atoms in the staples (c) are positively charged (+0.40e) reflecting their different oxidation state. The atomic charges in the Au_{13} icosahedron vertices are insensitive to the changes in environment.

 ISBN: 978-84-608-6082-2

It should be emphasized that the HOMO-LUMO gap and the optical absorption gap are not the same quantity. However, they are closely related and the reported HOMO-LUMO gaps for generalized gradient approximation (GGA) functionals display a strong correlation with the computed and/or experimental lowest optical transitions of AuNCs [24]. Furthermore, recent studies have shown for more accurate hybrid functionals that the lowest optical transition is over-estimated by the HOMO-LUMO gap [24,25].

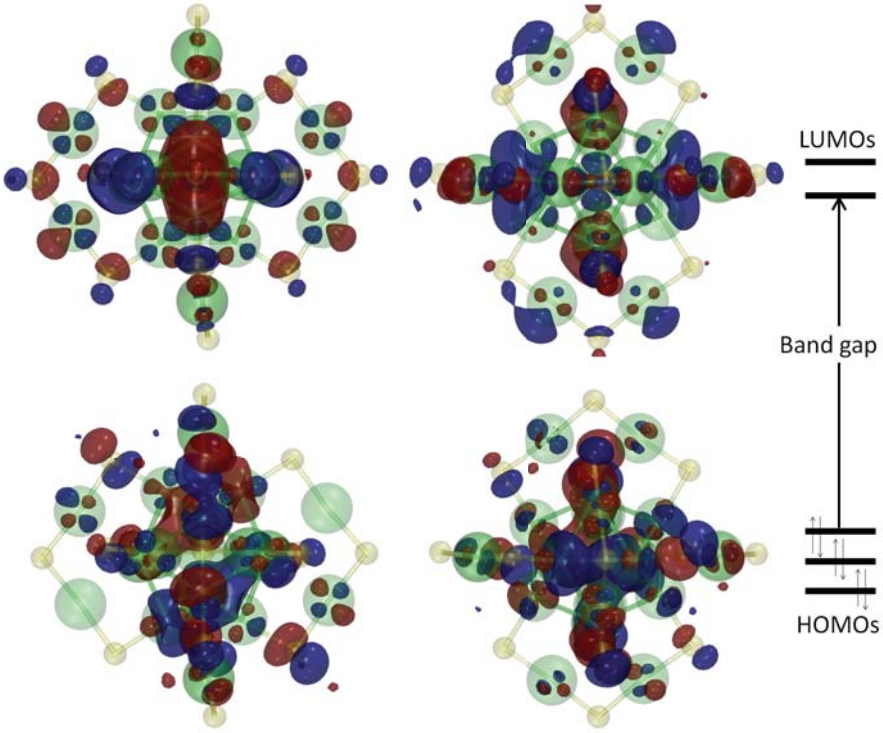


Figure 5. *Shapes of the Kohn-Sham orbitals for $Au_{25}(SR)_{18}^-$. HOMO (1P symmetry, bottom) and LUMO (1D symmetry, top) orbitals are displayed from two perspectives to better visualize their shapes, and they are depicted in blue and red to highlight different nodal signs. The cluster is shown transparent for clarity.*

The electronic structure of $Au_{25}(SR)_{18}^-$ corresponds to a closed-shell configuration of an 8-electron superatom with occupied 1S (two electrons) and 1P (six electrons, degenerate HOMO) orbitals, and the LUMO orbitals have 1D symmetry (Figure 5) [6,26]. The HOMO-LUMO gaps for the optimized structures of the studied systems are reported in Figure 4(B). The results clearly demonstrate how the solvent reduces the band gap significantly by 22% for the SCH_3 ligand system, with respect to the same system in vacuum (1.19 eV and 0.93 eV, respectively). However, this reduction by the solvent is in a wrong direction since the experimental optical gap is 1.3 eV [27].

 ISBN: 978-84-608-6082-2

Here, one should note that the HOMO-LUMO gap is expected to depend also on the ligand, as a consequence of the charge transfer effects and the variation in AuNC symmetry. In fact, replacing the ligand with GSH distorts the AuNC geometry, which adopts a C_1 symmetry instead of the original C_i symmetry of $Au_{25}(SCH_3)_{18}^-$. Most importantly, the HOMO-LUMO gap reduces further by 31% (0.93 eV vs. 0.65 eV), which means now an under-estimation of the experimental reference value by a factor of 2.

In general, hybrid exhange-correlation functionals such as PBE0 give more accurate results (larger values) for electronic band gaps than standard GGA functionals. The PBE0 value of $Au_{25}(GSH)_{18}^-$ is 1.54 eV and it is close to the measurement once the ligand is the same as in experiments [11]. Previously, it has been reported for AuNCs that PBE0 has poor performance for HOMO-LUMO gaps (severe over-estimation [24]), but we claim that this has been related to the limited description of the simulated system itself where the ligand has been different and solvent absent. For example, our PBE0 calculation gives here a value in the gas phase (2.20 eV), which is much too large and underlines the importance of including GSH ligands and solvent environment.

4. Conclusions

We have performed QM/MM simulations for the $Au_{25}(SR)_{18}^-$ clusters with two side groups to elucidate the effect of introducing a biological ligand (GSH) and explicit water solvent [12]. The overall changes in aqueous environment are consistent for both side groups although their characters are different, and this suggests that our results can be extended to other AuNCs. For atomic structure, our findings demonstrate that the icosahedral Au_{13} core and the six dimeric SR-Au-SR-Au-SR staple motifs are always preserved. However, the bond distances and angles exhibit visible changes under different conditions ("fluxionality"). The GSH ligands shield the metallic core and protect it towards nucleophilic attacks of external agents that can lead to undesired interactions. While methanethiolate (as other alkanethiols) is hydrophobic, the GSH ligands are hydrophilic and actively form hydrogen bonds with water and other neighboring ligands (amine and carboxylate groups). Changing the ligand from SCH_3 to GSH results in a different charge distribution which affects the HOMO-LUMO gap. Furthermore, GSH causes changes in the cluster symmetry due to the flexible nature of Au-Au bonds. This reduces the HOMO-LUMO gap further.

The HOMO-LUMO gap of AuNCs depends sensitively on the ligands and solvent environment. Previously, the role of solvent has been neglected in the theoretical studies of ligand-protected Au and Ag clusters, which have reported HOMO-LUMO gaps and/or optical absorption spectra (GGA) in good agreement with experiments [24,28]. However, we claim that this surprisingly good

 ISBN: 978-84-608-6082-2

correspondence (for standard DFT) is based on a cancellation of errors where the effect of missing solvent (and wrong ligands) is counterbalanced by the band gap underestimation.

Our QM/MM simulations show that the water solvent and ligand-exchange with GSH reduce the calculated (PBE) band gap drastically such that it underestimates the experiment by a factor of 2. However, once the hybrid PBE0 functional is used with QM/MM, we are able to produce a computed value (1.54 eV) that is slightly above the experimental optical gap. This result (HOMO-LUMO gap) presents an upper bound for the optical absorption gap and future work is needed to compute the optical absorption spectrum with the corresponding transitions.

5. References

[1] C.J. ACKERSON, M.T. SYKES, AND R.D. KORNBERG, *Defined DNA/Nanoparticle Conjugates*, Proc. Natl. Acad. Sci. **102** (2005) 13383-13385.

[2] C.J. ACKERSON, P.D. JADZINSKY, G.J. JENSEN, AND R.D. KORNBERG, *Rigid, Specific, and Discrete Gold Nanoparticle/Antibody Conjugates*, J. Am. Chem. Soc. **128** (2006) 2635-2640.

[3] M. ZHENG AND X. HUANG, *Nanoparticles Comprising a Mixed Monolayer for Specific Bindings with Biomolecules*, J. Am. Chem. Soc. **126** (2004) 12047-12054.

[4] P.D. JADZINSKY, G. CALERO, C.J. ACKERSON, D.A. BUSHNELL, AND R.D. KORNBERG, *Structure of a Thiol Monolayer-Protected Gold Nanoparticle at 1.1 Å Resolution*, Science **318** (2007) 430-433.

[5] M.W. HEAVEN, A. DASS, P.S. WHITE, K.M. HOLT, AND R.W. MURRAY, *Crystal Structure of the Gold Nanoparticle* $[N(C_8H_{17})_4][Au_{25}(SCH_2CH_2Ph)_{18}]$, J. Am. Chem. Soc. **130** (2008) 3754-3755.

[6] J. AKOLA, M. WALTER, R.L. WHETTEN, H. HÄKKINEN, AND H. GRÖNBECK, *On the Structure of Thiolate-Protected* Au_{25}, J. Am. Chem. Soc. **130** (2008) 3756-3757.

[7] Y. YU, Z. LUO, D.M. CHEVRIER, D.T. LEONG, P. ZHANG, D.-E. JIANG, AND J. XIE, *Identification of a Highly Luminescent* $Au_{22}(SG)_{18}$ *Nanocluster*, J. Am. Chem. Soc. **136** (2014) 1246-1249.

[8] T. DAINESE, S. ANTONELLO, J.A. GASCON, F. PAN, N.V. PERERA, M. RUZZI, A. VENZO, A. ZOLEO, K. RISSANEN, AND F. MARAN, $Au_{25}(SEt)_{18}$, *a Nearly Naked Thiolate-protected* Au_{25} *Cluster: Structural Analysis by Single Crystal X-ray Crystallography and Electron Nuclear Double Resonance*, ACS Nano **8** (2014) 3904-3912.

 ISBN: 978-84-608-6082-2

[9] T.W. NI, M.A. TOFANELLI, B.D. PHILLIPS, AND C.J. ACKERSON, *Structural Basis for Ligand Exchange on $Au_{25}(SR)_{18}$*, Inorg. Chem. **53** (2014) 6500-2.
[10] T. CAO, S. JIN, S. WANG, D. ZHANG, X. MENG, AND M. ZHU, *A Comparison of the Chiral Counterion, Solvent, and Ligand Used to Induce a Chiroptical Response from Au_{25}^- Nanoclusters*, Nanoscale **5** (2013) 7589-7595.
[11] Y. SHICHIBU, Y. NEGISHI, H. TSUNOYAMA, M. KANEHARA, T. TERANISHI, AND T. TSUKUDA, *Extremely High Stability of Glutathionate-protected Au_{25} Clusters Against Core Etching*, Small **3** (2007) 835-839.
[12] V. ROJAS-CERVELLERA, C. ROVIRA, AND J. AKOLA, *How do Water Solvent and Glutathione Ligands Affect the Structure and Electronic Properties of $Au_{25}(SR)_{18}^-$?*, J. Chem. Phys. Lett. **6** (2015) 3859-3865.
[13] J.C. PHILLIPS, R. BRAUN, W. WANG, J. GUMBART, E. TAJKHORSHID, E. VILLA, C. CHIPOT, R.D. SKEEL, L. KALE, AND K. SCHULTEN, *Scalable molecular dynamics with NAMD*, J Comput. Chem. **26** (2005) 1781-802.
[14] V. HORNAK AND C. SIMMERLING, *Generation of accurate protein loop conformations through low-barrier molecular dynamics*, Proteins **51** (2003) 577-590.
[15] W.L. JORGENSEN, J. CHANDRASEKHAR, J.D. MADURA, R.W. IMPEY, AND M.L. KLEIN, *Comparison of simple potential functions for simulating liquid water,* J. Chem. Phys. **79** (1983) 926-935.
[16] Y. ZHANG, T.-S. LEE, AND W. YANG, *A pseudobond approach to combining quantum mechanical and molecular mechanical methods*, J. Chem. Phys. **110** (1999) 46-54.
[17] A. LAIO, J. VANDEVONDELE, AND U. RÖTHLISBERGER, *A Hamiltonian Electrostatic Coupling Scheme for Hybrid Car–Parrinello Molecular Dynamics Simulations*, J. Chem. Phys. **116** (2002) 6941-6947.
[18] R. CAR AND M. PARRINELLO, *Unified approach for molecular dynamics and density-functional theory,* Phys. Rev. Lett. **55** (1985), 2471-2474.
[19] CPMD program, *Copyright IBM Corp. 1990-2015, Copyright MPI für Festkörperforscung, Stuttgart, 1997-2004.*
[20] S. NOSÉ, *A molecular dynamics method for simulations in the canonical ensemble*, Molecular Physics **52** (1984) 255-268.
[21] J.P. PERDEW, K. BURKE, AND M. ERNZERHOF, *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77** (1996) 3865-3868.
[22] C. ADAMO AND V. BARONE, *Toward reliable density functional methods without adjustable parameters: The PBE0 model*, J. Chem. Phys. **110** (1999) 6158-6170.
[23] A. VARGAS, G. SANTAROSSA, M. IANNUZZI, AND A. BAIKER, *Fluxionality of Gold Nanoparticles Investigated by Born-Oppenheimer Molecular Dynamics*, Phys. Rev. B **80** (2009) 195421.

 ISBN: 978-84-608-6082-2

[24] F. MUNIZ-MIRANDA, M.C. MENZIANI, AND A. PEDONE, *Assessment of Exchange-Correlation Functionals in Reproducing the Structure and Optical Gap of Organic-Protected Gold Nanoclusters,* J. Phys. Chem. C **118** (2014) 7532-7544

[25] J. ZHONG, X. TANG, J. TANG, J. SU, AND Y. PEI, *Density Functional Theory Studies on Structure, Ligand Exchange, and Optical Properties of Ligand-Protected Gold Nanoclusters: Thiolate versus Selenolate*, J. Phys. Chem. C **119** (2015) 9205-9214.

[26] M. WALTER, J. AKOLA, O. LOPEZ-ACEVEDO, P.D. JADZINSKY, G. CALERO, C.J. ACKERSON, R.L. WHETTEN, H. GRÖNBECK, AND H. HÄKKINEN, *A Unified View of Ligand-Protected Gold Clusters as Superatom Complexes*, Proc. Natl. Acad. Sci. **105** (2008) 9157-9162.

[27] S. LINK, A. BEEBY, S. FITZGERALD, M.A. EL-SAYED, T.G. SCHAAFF, AND R.L. WHETTEN, *Visible to Infrared Luminescence from a 28-Atom Gold Cluster,* J. Phys. Chem. B **106** (2002) 3410-3415.

[28] H. YANG, Y. WANG, H. HUANG, L. GELL, L. LEHTOVAARA, S. MALOLA, H. HÄKKINEN, AND N. ZHENG, *All-thiol-stabilized Ag_{44} and $Au_{12}Ag_{32}$ Nanoparticles with Single-crystal Structures*, Nat. Commun. **4** (2013) 2422.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

The nanofluidics of small droplets on hydrophobic surfaces

Alex Smith[1], Keoni Mahelona[1], and Shaun Hendy[2]

[1] *MacDiarmid Institute for Advanced Materials and Nanotechnology, Department of Physics, University of Auckland, Auckland 1010, New Zealand*

[2]*Te Pūnaha Matatini, Department of Physics, University of Auckland, Auckland 1010, New Zealand*

email: shaun.hendy@auckland.ac.nz

Abstract

The surfaces of many plant leaves are superhydrophobic, a property that may have evolved to help keep the leaves clean, by encouraging the beading and rolling of water droplets. Here we introduce effective slip into a model for a rolling droplet on a superhydrophobic surface and examine under what conditions slip might be important for the flow. In particular, we examine three limiting cases of the model where dissipation in the rolling droplet is dominated by viscous dissipation, surface friction or contact line friction respectively. We find that in molecular dynamics simulations of droplets on ideal surfaces, surface friction due to slip can dominate the motion of small nanoscale droplets, while in larger droplets motion is likely to be dominated by viscous shear, and slip can be neglected. On highly engineered superhydrophobic surfaces with large effective slip lengths, we find that contact line dissipation may play a role.

Key words: Superhydrophobic surfaces, droplets, effective slip

1. Introduction

Superhydrophobic surfaces that exploit the Lotus effect [1] to achieve contact angles close to 180^{o} are of interest both becuase of their role in biology [2], and for their potential technological applications [3]. Nature appears to use these structures to keep leaves clean: it is supposed that droplets are able to roll down the leaves, entraining dirt and contaminants as they go [4]. Indeed, experiments have found that droplets do roll rather than slide down superhydrophobic surfaces [5].

 ISBN: 978-84-608-6082-2

More recently, however, flows over superhydrophobic surfaces have been studied because they effectively violate the no-slip boundary condition [6]. When in the Cassie state, droplets or larger scale flows are essentially lubricated by a layer of air, leading to large effective slip lengths, with drag only occurring at the few points of the surface where the flow makes contact with the substrate. On such superhydrophobic surfaces effective slip lengths of micrometres to tens of micrometres have been observed [7, 8], scaling proportionally to the typical microstructural length scale [8, 9]. In some experiments, on highly ideal superhydrophobic surfaces, slip lengths of hundreds of microns have even been measured [10].

So do droplets slip as they roll down leaves? The leading model for the motion of droplets on superhydrophobic surfaces, due to Mahadevan and Pomeau [11], assumes a no-slip boundary condition [11], so in its current form it cannot be used to answer this question. In this talk we introduce slip into this model and examine under what conditions slip might be important for describing the flow. Recent experiments have found that on highly engineered superhydrophobic surfaces, rolling can be entirely suppressed [12]. Here, however, we will restrict ourselves to the case where droplets roll and slip, and examine the transition between these two regimes.

2. Results

We will consider droplets that are smaller in radius than the capillary length, $\kappa^{-1} = (\gamma/\varrho g)^{1/2}$, where γ is the droplet surface tension (for water $\simeq$ 2 mm). The contact zone radius, where the droplet is in contact with the surface, can be shown to be $\sim R^2/\kappa^{-1}$. The droplet is assumed to be on a titled superhydrophobic surface (at angle α to the horizontal), with a contact angle of 180°. The centre of mass velocity of the droplet, U, is considered to be the sum of its rolling velocity, U_r, plus a sliding velocity U_s due to slip at the droplet contact zone (see Fig. 1). In the contact zone, the velocity gradient $|\nabla u|$ goes as $\sim U_r/R$. An effective Navier slip boundary condition which is assumed to hold over the surface of the contact region relates the slip velocity to the velocity gradient in the contact zone: $U_s = b|\nabla u|$, so that $U_s = U_r\, b/R$ and $U \sim U_r(1+b/R) = U_s(1+R/b)$.

Here the slip length b should be considered an effective slip length [9]. The use of an effective slip length is justified provided the size of the contact surface, l, is much larger than the structuring of the superhydrophobic surface. For a surface composed of arrays of posts, with post spacing L, the effective slip length b is expected to scale as $L/\phi^{1/2}$ where ϕ is the area fraction of the surface covered by the posts. Thus we are justified in using an effective slip length provided $l \gg L \sim \phi^{1/2} b$.

 ISBN: 978-84-608-6082-2

In steady-state, the energy dissipated by the droplet must be balanced by the gain in gravitational potential energy. We consider two mechanisms for dissipation: viscous and frictional. The ratio of frictional dissipation to viscous dissipation can be shown to scale as b/l. If ($b/l << 1$) then viscous dissipation dominates then it can be shown that $U \sim (\gamma \kappa^{-1}/\mu R)(1+b/R)^2 \sin \alpha$, while if ($b/l >> 1$) then dissipation by frictional forces in the droplet contact zone is dominant and $U \sim (\gamma R/\mu b)(1+b/R)^2 \sin \alpha$. We test this theory using molecular dynamics simulations of the rolling droplet on superhydrophobic surfaces.

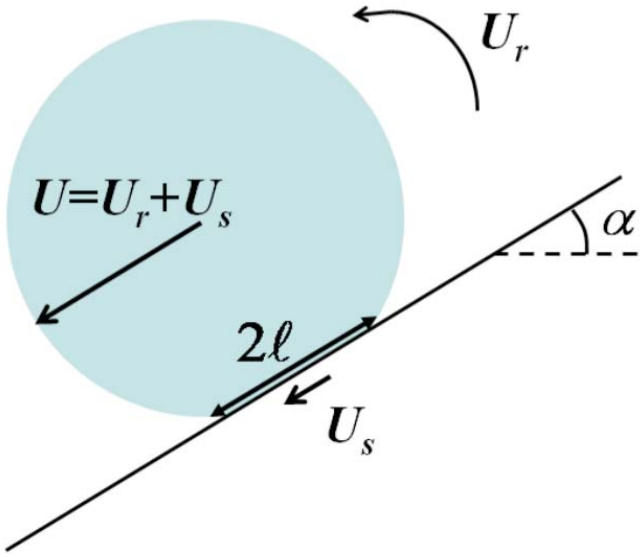


Figure 1: A droplet rolling down a superhydrophobic surface tilted at an angle α to the horizontal. The centre of mass velocity U is assumed to be the sum of a rolling component U_r and a sliding component U_s.

2. References

[1] W. BARTHLOTT AND C. NEINHUIS, PLANTA 202, 1 (1997).
[2] C. NEINHUIS AND W. BARTHLOTT, ANN BOT 79, 667 (1997).
[3] A. LAFUMA AND D. QUERE, NAT MATER 2, 457 (2003).
[4] D. QUERE, REPORTS ON PROGRESS IN PHYSICS 68, 2495 (2005).
[5] [5] D. RICHARD AND D. QUERE, EPL (EUROPHYSICS LETTERS) 48, 286 (1999).
[6] C. COTTIN-BIZONNE, J.-L. BARRAT, L. BOCQUET, AND E. CHARLAIX, NAT MATER 2, 237 (2003).
[7] C. H. CHOI AND C. J. KIM, PHYSICAL REVIEW LETTERS 96, 066001 (2006).

 ISBN: 978-84-608-6082-2

[8] P. JOSEPH, C. COTTIN-BIZONNE, J. M. BENOˆIT, C. YBERT, C. JOURNET, P. TABELING, AND L. BOCQUET, PHYSICAL REVIEW LETTERS 97, 156104 (2006).
[9] C. YBERT, C. BARETIN, C. C. BIZONNE, P. JOSEPH, AND L. BOCQUET, PHYS. FLUIDS 19, 123601 (2007).
[10] C. LEE, C. H. CHOI, AND C. J. C. KIM, PHYSICAL REVIEW LETTERS 101, 064501+ (2008).
[11] L. MAHADEVAN AND Y. POMEAU, PHYSICS OF FLUIDS 11, 2449 (1999).
[12] M. SAKAI, J.-H. SONG, N. YOSHIDA, S. SUZUKI, Y. KAMESHIMA, AND A. NAKAJIMA, LANGMUIR 22, 4906 (2006).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Ab initio modelling of semiconductor epitaxy processes - gas phase, surface and interfaces

Andreas Stegmüller[1], Phil Rosenow[1], Ralf Tonner[1]
[1] *Fachbereich Chemie and Material Sciences Centre, Philipps-Universität Marburg*

email: tonner@chemie.uni-marburg.de

Abstract

Ab initio methods have the ability to shed light on complex physical and chemical phenomena that are often hardly accessible experimentally. We employ quantum chemical computations, mainly density functional theory with dispersion correction, to understand elementary processes in different phases of metal organic vapour phase epitaxy (MOVPE), a technique that is used to grow highly ordered thin films of semiconductor materials e.g. for applications in optical devices. Analysis of chemical bonding and reactivity in the gas phase, on the surface and at the resulting interfaces of covalent semiconductor materials (e.g. silicon, gallium phosphide) helps to interpret experimental measurements and predicts improvements for this challenging procedure.
Elementary processes in all steps of MOVPE can have a crucial influence on the quality of the resulting films. In many cases, these key steps are hardly understood or experimentally inaccessible (e.g. reaction barriers) and theory can provide a crucial step forward in understanding.

Key words: density functional theory, growth modelling, ab initio, functional materials

1. Introduction

The controlled growth of thin films on surfaces is of high technological importance for the fabrication and performance optimization of semiconductor devices or optical coatings. One driving element is the need to move towards further miniaturization by a higher integration density of desired materials on a

 ISBN: 978-84-608-6082-2

smaller scale. To meet the need of advanced applications in the field of nanotechnology, reliable procedures for the controlled functionalization of surfaces have to be developed and understood. In contrast to "top-down" approaches imposing a defined structure onto a surface, the "bottom-up" approaches aim at building the desired features onto a surface by specific growth experiments. Prominent techniques to achieve the growth of films in a controlled fashion are the Metal Organic Vapour Phase Epitaxy (MOVPE or MOCVD) and Atomic Layer Deposition (ALD).
Elementary processes in all steps of these processes can have a crucial influence on the quality of the resulting films. In many cases, these key steps are hardly understood or experimentally inaccessible (e.g. reaction barriers). This provides an ideal ground for the application of ab initio approaches for the interpretation of experimental results and in a further step the prediction of material properties and the inspiration of new experiments. The goal is an improved fundamental understanding of these intriguing processes and the growth of materials with improved properties - e.g. less defects or better optical properties.

2. Quantum chemical approach to MOVPE

We employ quantum chemical computations, mainly density functional theory with dispersion correction, to shed light on the elementary processes in different phases of the MOVPE process. The state-of-the-art of this technique is depositing metastable materials (e.g. quaternary nitrides or bismides) which requires low growth temperatures. Therefore, chemical reactions in the gas phase and at the surface are the kinetic bottleneck for precursor decomposition and film growth.

The description of MOVPE processes by ab initio methods requires as a first step a detailed analysis of the possible gas phase reactions.[1] The first generation of MOVPE precursor molecules (hydrides) required high deposition temperatures and exhibited only limited gas phase chemistry.[2] The growth of metastable materials requires more information regarding the precursor chemistry since the low growth temperatures required imply complex gas phase decomposition chemistry of the precursors beyond simple homolytic cleavage steps.[3] For the gas phase chemistry of $Ga(C_2H_5)_3$ and $P(t\text{-}(C_4H_9)H_2)$ (known as TEGa and TBP in the MOVPE literature) we find very high barriers for several reactions advertised earlier[4] and unusual reaction mechanisms relevant for ligand design.[5, 6] The gas phase mechanism was also found to be relevant for the surface-assisted decomposition steps.

The surface termination plays a crucial role in setting up the right computational model for the investigations of surface adsorption and precursor reactivity. The transport of precursor molecules onto the heated substrate mostly uses hydrogen as a carrier gas. The adsorption and desorption kinetics of H_2 on Si(001) have been extensively covered in the experimental and theoretical literature.[7, 8] We

 ISBN: 978-84-608-6082-2

focus on the thermodynamically most stable surface termination taking into account the chemical potential of H_2 under MOVPE conditions (p = 50 mbar, T = 400 – 1200 °C) and find a good agreement with experimental findings only with elaborate computations of the full phonon spectrum.[9]

The epitaxial growth of a non-polar semiconductor material on a polar one (heteroepitaxy) results in formation of interfaces with interesting structural and electronic features. In a joint endeavour of experiment and theory, we could reveal the thermodynamic and kinetic reasons for the unusual appearance of GaP/Si interfaces. Intermixing models established previously are not sufficient to explain the pyramidal structure observed experimentally. Instead, relative facet stabilities in the interface region and adatom mobilities during growth need to be quantified to find a rationale for the atomistically-resolved experimental measurements.[10]

It is shown that quantum chemical approaches with and without periodic boundary conditions deliver detailed insight for a highly complex experimental procedure like metal organic vapour phase epitaxy.

3. References

[1] S. D. ELLIOTT, *Atomic-scale simulation of ALD chemistry,* Semicond. Sci. Technol. **27**, (2012), 074008.

[2] A. BRAUERS, *Alternative precursors for III-V MOVPE - promises and problems,* Prog. Cryst. Growth Charact. Mater. **22**, (1991), 1-18.

[3] G. B. STRINGFELLOW, *Fundamental aspects of organometallic vapor phase epitaxy,* Mater. Sci. Eng., B **87**, (2001), 97-116.

[4] A. STEGMÜLLER, P. ROSENOW, R. TONNER, *A quantum chemical study on gas phase decomposition pathways of triethylgallane (TEG, $Ga(C_2H_5)_3$) and tert-butylphosphine (TBP, $PH_2(t\text{-}C_4H_9)$) under MOVPE conditions,* Phys. Chem. Chem. Phys. **16**, (2014), 17018-17029.

[5] A. STEGMÜLLER, R. TONNER, *A quantum-chemical descriptor for CVD precursor design: predicting decomposition rates of TBP and TBAs isomers and derivatives,* Chem. Vap. Deposition **21**, (2015), 161-165.

[6] A. STEGMÜLLER, R. TONNER, *β-Hydrogen Elimination Mechanism in the Absence of Low-Lying Acceptor Orbitals in $EH_2(t\text{-}C_4H_9)$ (E = N–Bi),* Inorg. Chem. **54**, (2015), 6363-6372.

[7] A. Groß, *Theoretical Surface Science - A Microscopic Perspective*, Springer, Berlin, Heidelberg, **2009**.

[8] M. DÜRR, U. HÖFER, *Dissociative adsorption of molecular hydrogen on silicon surfaces,* Surf. Sci. Rep. **61**, (2006), 465-526.

[9] P. ROSENOW, R. TONNER, *Extent of hydrogen coverage of Si(001) under chemical vapor deposition conditions from ab initio approaches,* arxiv: 1603.02918.

 ISBN: 978-84-608-6082-2

[10] A. BEYER, A. STEGMÜLLER, J. O. OELERICH, K. JANDIERI, K. WERNER, G. METTE, W. STOLZ, S. BARANOVSKI, R. TONNER, K. VOLZ, *Pyramidal Structure Formation at the Interface between III/V Semiconductors and Silicon,* Chem. Mat. DOI: 10.1021/acs.chemmater.5b04896, (2016).

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Essential Collective Dynamics of Biological Polymers

Maria Stepanova[1,2]

[1] *Department of Electrical and Computer Engineering,*
University of Alberta;
[2] *Department of Physics, Astronomy, and Materials Science,*
Missouri State University

emails: ms1@ualberta.ca, mariastepanova@missouristate.edu

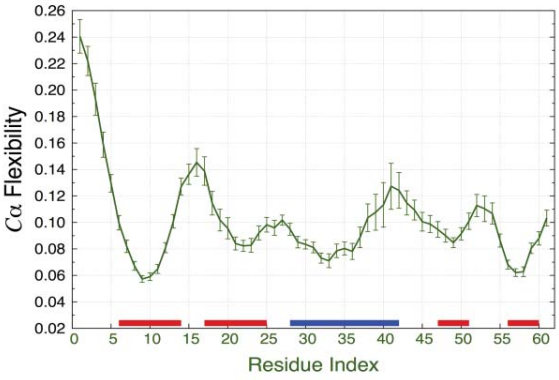
Abstract

Key words: biomolecular dynamics, structural biology software

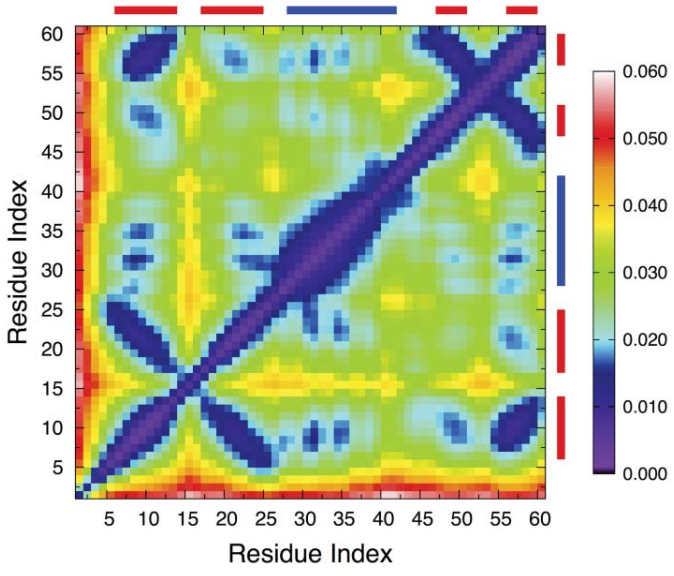
Recent advances in obtaining thousands of three-dimensional structures of proteins and nucleic acids, complemented by atomic-resolution molecular dynamics (MD) simulations, have allowed researchers to see motion of biopolymers in atomic details. However drawing accurate, specific, and quantitative predictions of how biopolymers respond to various stimuli remains a challenge. The difficulty is that biopolymers exhibit intricate multidimensional dynamics that involve a multitude of timescales, from fast picosecond-range thermal motions of small atomic groups to slow large-scale concerted motions spanning from nanoseconds to micro-second regimes and longer. Such slow motions are believed to determine the cellular functions of biopolymers in the first place. Over the recent years, we developed a novel computational framework, which we denote as the essential collective dynamics (ECD) analysis, allowing to accurately characterizing the dynamics of slow collective motions in macromolecules [1-10]. The framework stems from our statistical-mechanical analysis of generalized Langevin dynamics [1,3,4], allowing to identify invariant (stable and persistent) correlations of atomic groups' motion from short fragments of MD simulations of macromolecules in solution. Subsequently, we have derived and tested a suite of specific descriptors, such as dynamics domains [1-5], main-chain and side-chain pair correlations [4,5], and main-chain flexibilities [2-6]. The predictions that we have obtained from sub-nanosecond or nanosecond-scale MD trajectories agree very well with X-ray [8] and NMR [1-5,7] structural data, some of which represent significantly longer timescales. The ECD framework has proved to be very efficient in characterizing dynamical stability of individual protein molecules [1,2,5,7-9], as well as protein-protein [5,8,9], protein-DNA [6], protein-small molecule [8,9], and protein-nanoparticle [10] complexes.

 ISBN: 978-84-608-6082-2

This presentation introduces our software tool for ECD analysis of biopolymers, which we are preparing for open-source release [11]. The tool is implemented as a Linux package comprising a set of C++ applications integrated into a Python wrapper. Using short sub-nanosecond segments of production MD trajectories, pre-built scripts allow doing accurate predictions about main-chain flexibility profiles and/or pair correlations in a macromolecule. In this framework, exhaustive conformational sampling is avoided by addressing invariant dynamical modalities [4], which would pertain beyond the sampling data set.



(A)



(B)

Figure 1: Examples of ECD analysis for protein G (Protein Data Bank ID 1IGD). (A) – main-chain fexibility profile; (B) – main-chain pair correlations map [11]. The blue bars represent α-helix and red bars represent β-strands. In panel (B), low levels of the descriptor (purple and blue colors) identify stronger pair correlations.

 ISBN: 978-84-608-6082-2

Acknowledgements:

The author is happy to thank Nikolay Blinov, Mark Berjanskii, Santo Poulose, Bilkiss Issack, Taras Fito, Oliver Stueker, Lyudmyla Dorosh, and Jonathan Mane, whose dedicated work and extensive programming contributions made it possible to develop and test the new software tool.

References:

[1] M. STEPANOVA, *Dynamics of essential collective motions in proteins*, Phys. Rev. E **76** (2007) 051918.
[2] N. BLINOV, M. BERJANSKII, D. S. WISHART, AND M. STEPANOVA, *Structural domains and main-chain flexibility in prion proteins*, Biochemistry **48** (2009) 1488–1497.
[3] M. STEPANOVA, *Identification of dynamic structural domains in proteins, analysis of local bond flexibility, and application for interpretation of NMR experiments*, Molecular Simulation, **37** (2011) 729-732.
[4] A. POTAPOV AND M. STEPANOVA, *Conformational modes in biomolecules: Dynamics and approximate invariance*, Phys.Rev.E **85** (2012) 020901 (R).
[5] B.B. ISSACK, M. BERJANSKII, D.S. WISHART, AND M. STEPANOVA, *Exploring the essential collective dynamics of interacting proteins: application to prion protein dimers*, PROTEINS: Structure, Function, and Bioinformatics **80** (2012) 1847-1865.
[6] K. BARAKAT, B. B. ISSACK, M. STEPANOVA, AND J. TUSZYNSKI, *Comparative analysis of essential collective dynamics and observed NMR flexibility profiles in evolutionarily diverse prion proteins*, Prion **5** (2011) 188 – 200.
[7] K.P. SANTO, M. BERJANSKII, D.S. WISHART, AND M. STEPANOVA, *Effects of temperature on the p53-DNA binding interactions and their dynamical behaviour: Comparing the wild-type to the R248Q mutant*, PLoS ONE **6** (2011) e27651.
[8] L. DOROSH, O.A. KHARENKO, N. RAJAGOPALAN, M.C. LOEWEN, AND M. STEPANOVA, *Molecular mechanisms in the activation of abscisic acid receptor PYR1*, PLoS Computational Biology **9** (2013) 1003114.
[9] L. DOROSH, N. RAJAGOPALAN, M.C. LOEWEN, AND M. STEPANOVA, *Molecular mechanisms in the selective basal activation of pyrabactin receptor 1: Comparative analysis of mutants*, FEBS Open Bio **4** (2014) 496-509.
[10] O. STUEKER, V. ORTEGA, G. GOSS, AND M. STEPANOVA, *Understanding interactions of functionalized nanoparticles with proteins: A case study on lactate dehydrogenase*, Small **10** (2014) 2006-2014.
[11] J. MANE AND M. STEPANOVA, *Essential collective dynamics analysis user guide, Version 1.0*, University of Alberta, Edmonton, 2016.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Modelling of Nanoparticle-Enzyme Complex

Oliver Stueker[1] and Maria Stepanova[1,2]

[1] *Department of Electrical and Computer Engineering,*
University of Alberta;
[2] *Department of Physics, Astronomy, and Materials Science,*
Missouri State University

emails: ostueker@gmail.com, ms1@ualberta.ca

Abstract

Key words: nano-biological conjugates, molecular simulation

Nano-bio-electro-mechanical engineering is a rapidly advancing interdisciplinary field of research and innovation. Emerging technologies integrating biological components with nanostructured solid surfaces are expected to revolutionize biomedical instrumentation, environmental diagnostics, and green energy harvesting capabilities. The ongoing transformative convergence of information and biomedical technologies requires, in the first place, efficient bio-sensors and bio-actuators allowing for adaptive and addressable interfacing of electronic and biological components. However, in order to design such systems rationally, computational studies are required that would allow to better understand and predict the properties of all parts of the integrated systems. The challenge is that, traditionally, modeling of solid state materials and biological polymers has been largely unrelated. In order to efficiently predict properties of nano-biological conjugates, cross-disciplinary modeling approaches are required. In our recent work [1] we developed a detailed, all-atom platform for predicting the interaction of proteins with bio-functionalized metallic surfaces in solution. As an example, we have investigated a binding of L-lactate dehydrogenase (LDH) enzyme with gold nanoparticle (NP) decorated with a monolayer of mercapto-undecanoic acid (MUA). This presentation addresses challenges of such modelling, as well as our suggested solutions. An atomistic molecular mechanics model has been used to describe a gold nanoparticle with a diameter of ~4.2 nm decorated with 386 MUA molecules (Figure 1). The model is based on the OPLS-AA force field [2] with proper extensions [3,4]. The model of the MUA monolayer has been validated by simulating a similar monolayer on a planar, yet atomistic gold surface (Figure 2), and comparing the results with existing published works [5]. Molecular dynamics (MD) simulations have been conducted [1] for the three systems comprising free

 ISBN: 978-84-608-6082-2

LDH enzyme, functionalized gold NP, and a NP-LDH complex in water (Figure 3), for 10 ns each at T=300K using the GROMACS software package. The orientation of LDH in the complex with the NP was chosen according to the respective electrostatic potentials.

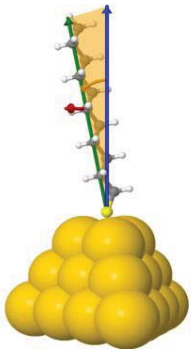


Figure 1: A single alkyl-thiol molecule on a gold surface.

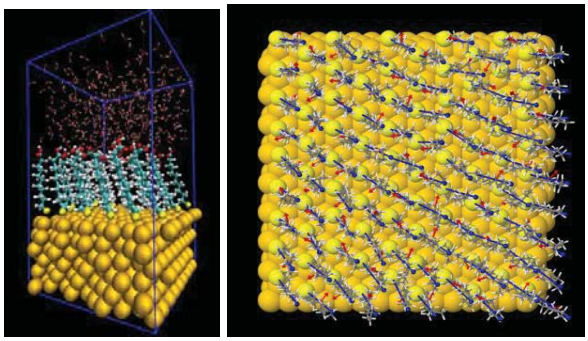


Figure 2: MUA monolayer on a planar atomistic gold surface, constructed for benchmark validation of the model of decorated NP surface.

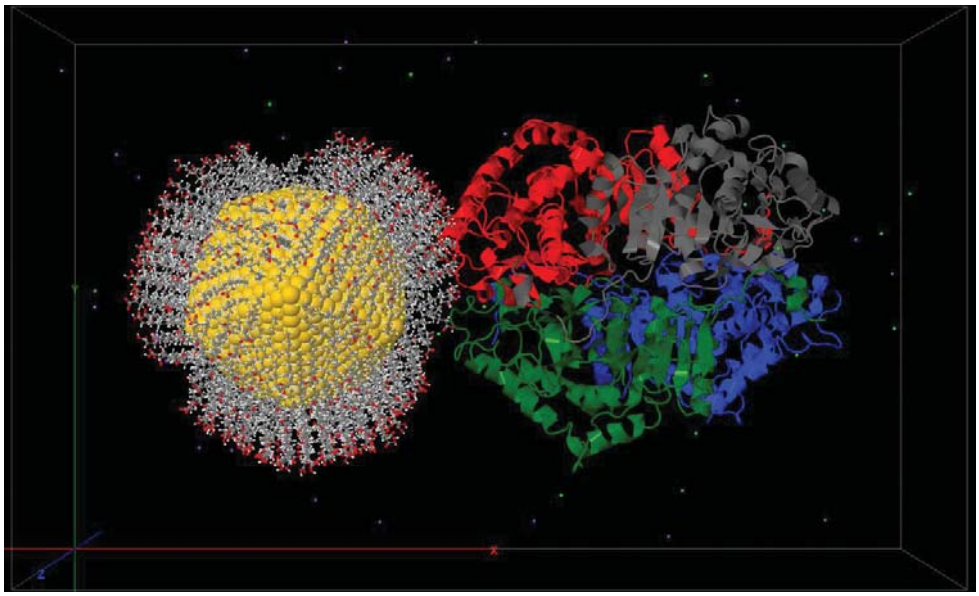
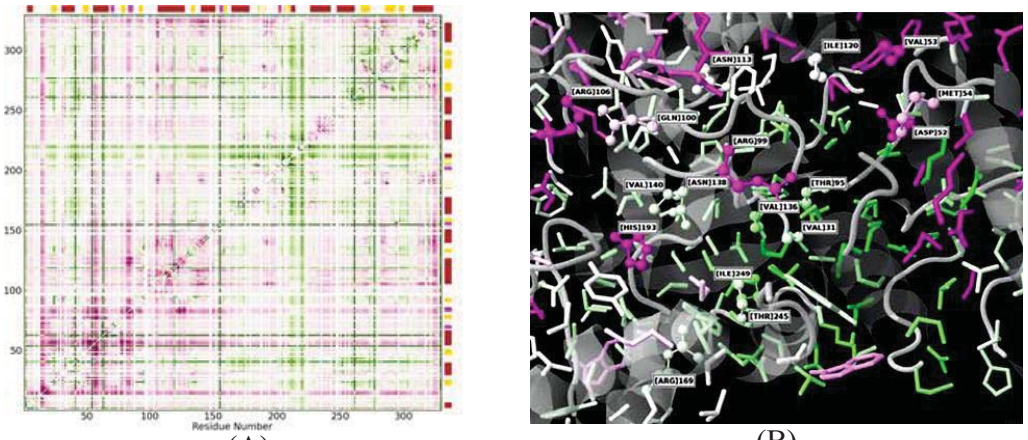


Figure 3: Snapshot of MD simulation for the nanoparticle-LDH complex [1]. In the LDH gray, red, green, and blue colors show the four identical subunits.

Employing our original essential collective dynamics (ECD) framework, we have investigated how the binding of NP influences the conformation of LDH tetramer structure (Figure 4). The results demonstrate that, although the dynamics of LDH

 ISBN: 978-84-608-6082-2

main chains exhibit only a minor response to the NP binding, the dynamics of side chains are significantly altered in all four active sites of the enzyme. These results also demonstrate that atomistic modeling of NP-protein interactions, complemented by the essential collective dynamics analysis, can be used to understand the interaction of biopolymers with nanoparticles, and potentially other nanostructured objects.



(A) (B)

Figure 4: Change of intra-molecular side-chain correlations in LDH upon the NP binding. (A) – intramolecular difference correlation maps of side chains in one of the LDH subunits; (B) – the change of correlations color-mapped onto the area of active site in the subunit. The magenta color indicates areas which become more constrained, and green color shows less constrained areas in comparison to the free LDH. The red, yellow and pink bars represent α-helices, β-sheets, and 3/10-helices, respectively.

References:

[1] O. STUEKER, V. ORTEGA, G. GOSS, AND M. STEPANOVA, *Understanding interactions of functionalized nanoparticles with proteins: A case study on lactate dehydrogenase*, Small **10** (2014) 2006-2014.

[2] W.L. JORGENSEN AND J. TIRADO-RIVES, *Potential energy functions for atomic-level simulations of water and organic and biomolecular systems*, Proc.Natl Acad Sci. USA **102** (2005) 6665-6670.

[3] F. IORI, R. DI FELICE, E. MOLINARI AND S. CORNI, *GolP: An atomistic force-field to describe the interaction of proteins with Au(111) surfaces in water,* J. Comput. Chem. **30** (2009) 1465-1476.

[4] R. POOL, P SCHAPOTSCHNIKOW AND T.J.H. VLUGT, *Solvent effects in the adsorption of alkyl thiols on gold structures: A molecular simulation study,* J. Phys. Chem. C, **111** (2007) 10201-10212.

[5] J. HAUTMAN AND M.L. KLEIN, *Simulation of a monolayer of alkyl thiol chains,* J. Chem. Phys **91** (1989) 4994-5001.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

Comparison of data mining tools

Antonio J. Tallón-Ballesteros[1] and Jonathan E. Benavides-Vallejo[2]

[1] *Department of Languages and Computer Systems, University of Seville, 41012 Spain*
[2] *Higher Technical School of Computer Science Engineering, University of Seville, 41012 Spain*

emails: atallon@lsi.us.es, jonbenval1@alum.us.es

Abstract

This paper introduces a comparison between two machine learning tools such as Weka and Rapid Miner. We describe the functionalities of both tools and also the installation and running procedure. An analysis tab by tab or menu by menu is carried out in order to state the main task that are available in any platform.

Key words: machine learning, tools, Weka, Rapid Miner, data mining, comparison

1. Introduction

Nowadays, the variety of data mining tools is very wide. Some tools like R, Weka [1], Rapid Miner [2], Orange and Excel Miner are without any doubt the most popular software packages to analyse data and to discover knowledge in the raw data. Weka is a very well-known platform that has been developed under the Java programming language. Currently, the usage of Rapid Miner is making progress. One key point is that their interface is very similar to the graphical design products and the user could create a sheet that will populated with drag and drop actions. Their programing language is also Java.

This paper describes Weka and Rapid Miner tools. In both case a Java Virtual Machine (JVM) is required.

 ISBN: 978-84-608-6082-2

2. Weka

The installation of Weka is very easy. We can use the installer or even with can directly download the .jar file and we can directly run weka.jar if we have one JVM installed. It is recommendable to run it under JDK 1.7 or 1.8 to be able to use the latest versions of Weka that incorporate new functionalities. Once we have started up the application the following screen will appear:

Figure 1.- Initial screen of Weka.

The top menu offer four options and among them Visualisation and Tools are related, respectively, with graphs (visualizers or plots) and viewers to see the data in plain arff or SQL format. Arff is the native format for Weka data sets. The frame application depicts the main modules of Weka.

Explorer is likely the most executed part all over the world and is also the suitable place for new users to start the "learning" and to deepen into the software. After clicking on Explorer the following window will have the focus:

 ISBN: 978-84-608-6082-2

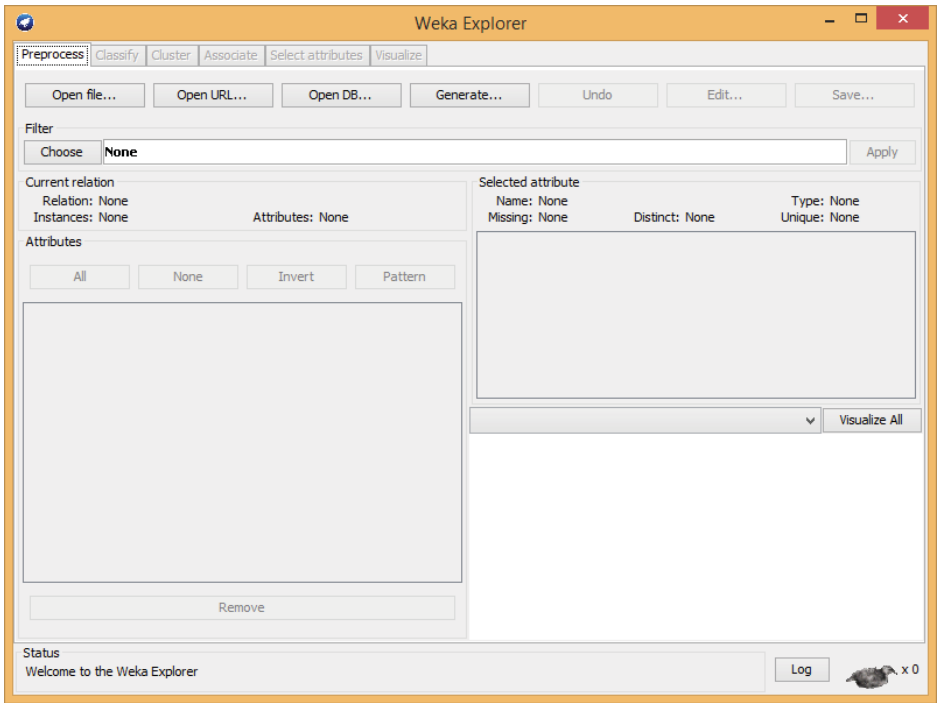


Figure 2.- Screen of Weka Explorer.

Explorer is a visual environment that offers a Graphical User Interface (GUI) to access to all the packages. It is very outstanding to mention that only the option that could be run at every moment are available to the user. There are six tabs. The first tab called Preprocess and is ready to load a data set in different formats such as .arff, .csv, .bsi or .xrff. After the data loading the remaining uppers tabs will be enabled and also the Filter frame of the current tab is able to be executed whose main purpose is to transform the data at the instance of feature level. The second tab named Classify is designed to execute a classifier with using the loaded file (automatically will be created the training and test sets) or even providing an additional file that could be the test set. The third tab offers some clusterers and their interface is very similar to "Classify". The fourth tab is simpler and the user could specify the associator and the start it or in the meanwhile top stop it. The tab Select attributes provides different implementation of some attribute selection algorithms. Finally, the tab Visualize offer diffent kind of plots.

Experimenter is a module to configure experiments and data analysis with different data files.

KnowledgeFlow is an application to create machine learning projects by means of flow diagrams.

Simple CLI is a text-mode client to access to Weka packages with commands.

 ISBN: 978-84-608-6082-2

3. Rapid Miner

The installation of Rapid Miner should be done only using the installer. Figure 3 shows the screen that initially appears and is similar to any office suite application with the File, Edit, View and Help menus, among other. The first step that we need to do is a Create a New Process. Once we chose this action a new windows, as depicted in Figure 4, will receive the focus. A process is the basic element is Rapid Miner and is similar to a sheet where the user should define their design using the Operators (objects) that are available.

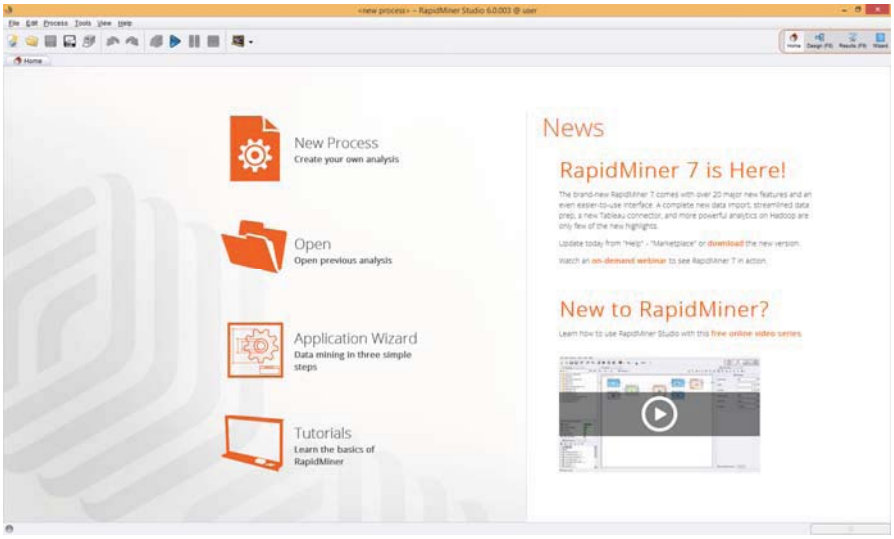


Figure 3.- Initial screen of Rapid Miner.

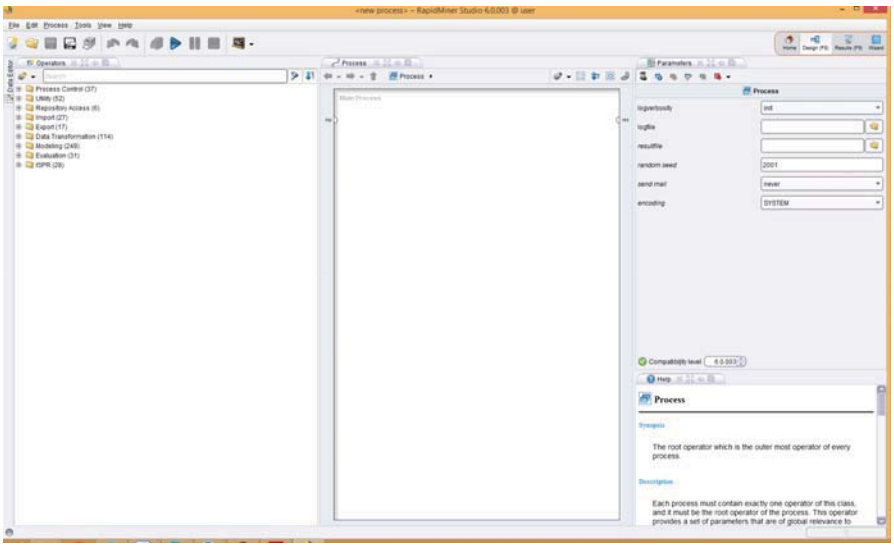


Figure 4.- New process screen of Rapid Miner.

 ISBN: 978-84-608-6082-2

In the left part we can see the view Operators. In the mid of the windows there is a sheet named Main Process which is the place to Drag and Drop the desired Operators and to connect them. The right part provides the current values of the parameters and also the compatibility level. The folders on the right are all the packages that could be deployed in order to use a concrete operator. We can use the Repository Access to select the operator Retrieve in order to load a data set or even the Import folder to load an external data set. Data Transformation is similar to the Preprocess tab of Weka. The folder Modeling contains all the typical tasks of Data Mining such as Classification and Regression (inside is include also Rule Induction related with Subgroup Discovery), Attribute Weighting, Clustering and Association.

To conclude this section, Figure 5 shows an example of a design ready to run about a classification task with a decision tree using and training and test set.

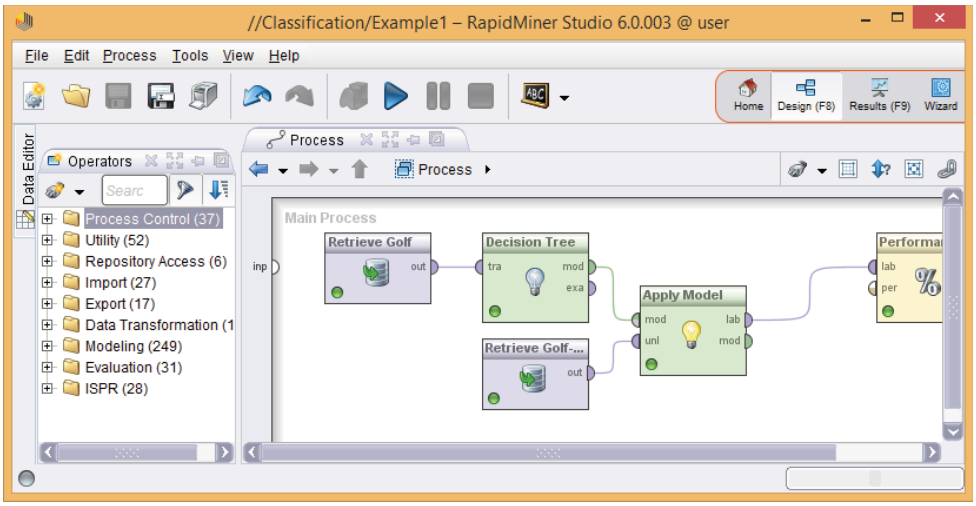


Figure 5.- Classification process in Rapid Miner.

4. Conclusions

Both tools offer a great deal of possibilities to carry out any kind of data mining tool. Learning about Weka may be easier because the user does not need to design the workflow and only the available option could be chosen and the different kind of machine learning actions are separated in different tabs. On the other hand, Rapid Miner is mainly a visual tool where the user may find difficulties at the beginning because the operators may not be very simple to master or even to find which is the name of the operator because the number of them may be in the order of four thousands. The appearance of Rapid Miner is similar to other kind of

 ISBN: 978-84-608-6082-2

computer science tools such and Integrated Development Environments for languages like C++, Java or web languages.

5. References

[1] M. HALL et al., *The WEKA data mining software: an update*. ACM SIGKDD explorations newsletter **11(1)** (2009) 10-18.
[2] I. MIERSWA, *Rapid Miner*. KI **23(2)** (2009) 62-63.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4-8 July, 2016

On two optimisation problems related to unsatisfied demand on a time interval

M.T.Todinov
Department of Mechanical Engineering and Mathematical Sciences
Oxford Brookes University, Oxford, OX33 1HX
email: mtodinov@brookes.ac.uk

Abstract

This paper focuses on two important optimisation problems: (i) the maximum size of the system that can be serviced by a given number of sources so that the unsatisfied demand does not exceed a tolerable level and (ii) the minimum number of sources needed to service random demands so that the unsatisfied demand does not exceed a tolerable level. To solve these problems, a computational framework for determining the expected fraction of unsatisfied demand on a time interval has been created and closed-form solutions for the expected fraction of unsatisfied demand have been derived.

Key words: unsatisfied demand, constraint on resources, probability, risk, random demand, sources, consumers

1. Introduction

Risk is often linked with the collision of random demands for a particular resource on a finite time interval. Unsatisfied demand occurs if one or more demands arrive at a time during which all available sources are engaged in servicing other demands.

The need for assessing this risk is often present in many manufacturing processes where a number of machine centres demand expensive measuring equipment, control equipment, production equipment or an operator, at a random time during the production process. Because the control equipment is expensive and unique, it is usually not feasible to equip each machine centre with a separate piece of equipment. The demands for the resource may occur at random times during a shift. If the start of demand for consumer i is denoted by 'si' and the end of demand by 'ei', Fig.1 illustrates the problem by three machine centres demanding X-ray portable equipment for measuring the residual stresses at the surface of quenched components, at random times $s1$,$s2$ and $s3$, for duration intervals (s1,e1), (s2,e2) and (s3,e3).

 ISBN: 978-84-608-6082-2

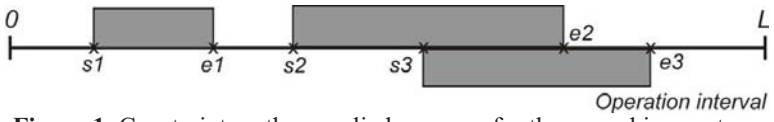


Figure 1. Constraint on the supplied resource for three machine centres.

If a single piece of control equipment is available, a simultaneous demand from more than a single centre cannot be satisfied. The overlapping region (s3, e2) in Fig.1 marks unsatisfied demand on the supplied resource.

For a complex systems (production system, computer network, etc.), the random demands are often failures of the separate components building the system each of which demands a repair resource.

In queuing theory, the Poisson process has been traditionally used as a statistical model for random events occurring in a time interval (Khintchine, 1969; Gross & Harris, 1985; Gnedenko & Kovalenko 1989; Thompson, 1988; Allen 1990). Neither of these classical comprehensive texts, nor more recent comprehensive texts devoted to various problems in queuing theory and probability (Kalashnikov 1994; Giambene 2005; Weiss, 2006) treats the question related to risk of unsatisfied demand from random requests on a time interval.

From Figure 1, it is clear that the problem of unsatisfied demand can be reduced to a problem of geometrical probability where a segment of specified length L is covered by randomly placed smaller segments with different lengths. The probability of unsatisfied demand can then be estimated by the probability of an overlap by two or more than two segments. For a single available source, the expected time of unsatisfied demand is numerically equal to the expected fraction of area covered simultaneously by two or more than two segments.

There are already a number of publications related to covering the circumference of a circle with segments or a linear segment with segments (Solomon 1978; Stevens 1939; Shepp 1972; Coffman et al., 1998; Coffman et al., 1994; Justicz et al. 1990).

Stevens (1939) derived closed-form expressions for the probability of covering a circle by a specified number of randomly positioned arcs of the same length and for the probability of existence of a specified number of uncovered gaps. Randomly positioned arcs with different lengths have been considered in Shepp 1972, where a condition has been derived for covering the circle with probability equal to one. A segment covered by randomly positioned smaller segments has been considered in (Justicz et al. 1990) where the probability of existence of a segment which intersects every other segment has been estimated.

Despite the substantial progress made in problems related to covering a large segment with randomly positioned smaller segments, no study seems to exist related to estimating the expected lineal fraction covered by m or more random segments. The answer to this question however, is the key to evaluating the risk of unsatisfied demand.

An important aspect of the problem related to risk associated with unsatisfied demand on a time interval is the maximum tolerable level of the probability of

unsatisfied demand. The tolerable level of the probability of unsatisfied demand depends on the magnitude of the consequences resulting from unsatisfied demand and must be set individually, by the risk experts in the respective application area. Thus, for injured or critically ill patients demanding life-saving medical equipment, the consequences of unsatisfied demand can be grave. In this case, unsatisfied demand means human fatalities and the maximum tolerable level of the probability of unsatisfied demand is very low. In the case where the monitoring services of a single available operator are needed for the successful operation of a number of machines, each of which places a random demand, the consequences of unsatisfied demand are significant and the maximum tolerable level of the probability of unsatisfied demand is low. For machine centres demanding for example measuring equipment to control the surface roughness of machined work-pieces, the consequences of unsatisfied demand are moderate and the corresponding tolerable level of the probability of unsatisfied demand is moderate.

Questions of significant practical importance, directly related to the problem of unsatisfied random demand, are:

(i) What is the minimum number of sources servicing a given number of consumers so that the probability of unsatisfied demand remains below the maximum tolerable level?

(ii) What is the maximum number of consumers that can be serviced by a single source so that the probability of unsatisfied demand remains below the maximum tolerable level?

The purpose of this study is to provide answers to these questions. The answer is important to find the optimal balance between the number of provided sources and the risk of unsatisfied demand. More provided sources than the optimal number is costly and undermines the profitability of the enterprise; fewer provided sources than the optimal number increases the risk of unsatisfied demand which leads to a disrupted production

2. Estimating the risk of unsatisfied demand by the expected fraction of the time of unsatisfied demand

Suppose that m pieces of a particular resource are available which can satisfy m simultaneous demands, but not m+1 or more simultaneous demands. If the different demands are represented as overlapping segments with different lengths d_i, the risk of unsatisfied demand can also be estimated with the expected fraction of length covered simultaneously by more than m random segments randomly placed along a segment with length L.

Consider a case where the duration of the demand for the ith consumer is equal to d_i, during the operation period with length L. The ratio of the duration of the demand and the time interval 'L' will be denoted by $\psi_i = d_i / L$.

Before determining the expected time fraction of unsatisfied demand, the following theorem related to a coverage of space with volume V by n 3-D

 ISBN: 978-84-608-6082-2

randomly placed interpenetrating objects with volumes v_i $(i=1,...,n)$, will be stated and proved. The volume fractions of the separate objects will be denoted by $\psi_i = v_i / V$. The coverage of a point from the volume V is a 'coverage of order k' if exactly k objects cover the point. The following theorem then holds.

Theorem 1. *The expected covered fraction of order k (k=0,1,...,n) by n interpenetrating objects with volume fractions* ψ_i*, is given by the k+1st term of the expansion* $\prod_{i=1}^{n}[(1-\psi_i)+\psi_i]$.

Proof. The volume fraction covered by exactly m objects can be determined from the probability that a randomly selected point in the volume V will sample simultaneously m overlapping (interpenetrating) random objects. The probability that a randomly selected point in the volume V will sample simultaneously m overlapping objects is equal to the probability that a fixed point from the volume V will be covered exactly m times by randomly placed objects in the volume V.

Let p_0 denote the probability that the fixed point will not be covered; p_1 denote the probability that the fixed point will be covered by exactly one random object,...,and p_n denote the probability that the fixed point will be covered by all n random objects.

Because the locations of the random objects are statistically independent events, the probability of the event A_0 that a fixed point in the volume will not be covered by any of the random objects given by

$$P(A_0) = \prod_{i=1}^{n}(1-\psi_i)^n, \tag{1}$$

which is the probability that the fixed point will not be covered by the first, the second,...,the nth object.

The probability of the event A_1 that exactly one random object will cover the fixed point is a sum of the probabilities of the following mutually exclusive events: the first object covers the fixed point and the rest of the random objects do not; the second object covers the fixed point and the rest of the objects do not and so on. As a result, the probability $P(A_1)$ that the fixed point will be covered by exactly one random object is given by

$$P(A_1) = \sum_{i=1}^{n}\left(\psi_i \prod_{\substack{k=1 \\ k\neq i}}^{n}(1-\psi_k)\right) \tag{2}$$

The probability that exactly two random objects will cover the fixed point is a sum of the probabilities of the following mutually exclusive events: the first and the second random object cover the fixed point and the rest of the objects do not; the first and the third random object cover the fixed point and the rest of the random objects do not and so on, until all possible combination of two objects out

 ISBN: 978-84-608-6082-2

of *n* are exhausted. As a result, the probability that the fixed point will be covered by exactly two random objects is given by

$$P(A_2)=\sum_{i1,i2)}\left(\psi_{i1}\psi_{i2}\prod_{\substack{k=1\\k\neq i1;k\neq i2}}^{n}(1-\psi_k)\right) \tag{3}$$

where $\sum_{i1,i2}$ denotes the sum over all possible combinations of two indices *i1* and *i2* out of *n*. The number of these combinations is $\binom{n}{2}=\frac{n!}{2!(n-2)!}$.

Continuing this reasoning through the cases 3,4,...,*n*, the probability $P(A_m)$ that the fixed point will be covered by exactly *m* random objects is given by

$$P(A_m)=\sum_{i1,\ldots jm}\left(\psi_{i1}\psi_{i2}\ldots\psi_{im}\prod_{\substack{k=1\\k\neq i1;\ldots k\neq im}}^{n}(1-\psi_k)\right) \tag{4}$$

where $\sum_{i1,\ldots jm}$ denotes the sum over all distinct combinations of *m* indices *i1,i2,...,im* out of *n*. The number of these combinations is $\binom{n}{m}=\frac{n!}{m!(n-m)!}$.

The fixed point can either remain uncovered, covered by exactly one, two,...,*n* objects and there are no other alternatives. Therefore, the events $A_0, A_1, \ldots, A_n$ are mutually exclusive and exhaustive. According to the third axiom of the probability theory, their probabilities add up to one:

$$\sum_{i=0}^{n}P(A_i)=1 \tag{5}$$

$$\prod_{i=1}^{n}(1-\psi_i)^n+\ldots+\sum_{i1,\ldots jm}\left(\psi_{i1}\psi_{i2}\ldots\psi_{im}\prod_{\substack{k=1\\k\neq i1;\ldots k\neq im}}^{n}(1-\psi_k)\right)+\ldots+\prod_{i=1}^{n}\psi_i{}^n=1 \tag{6}$$

Equation (6) can also be presented as an expansion of the expression $\prod_{i=1}^{n}[(1-\psi_i)+\psi_i]$. The theorem has been proved.■

Theorem 2. *The expected fraction of order k (k=0,1,...,n) from the volume V, covered by n interpenetrating objects with volume fractions* ψ_i*, coming from a particular distribution with mean* $\overline{\psi}$ *is given by the k+1st term of the expansion* $[(1-\overline{\psi})+\overline{\psi}]^n=\sum_{i=0}^{n}\binom{n}{i}\overline{\psi}^i(1-\overline{\psi})^{n-i}=1$.

Proof.

 ISBN: 978-84-608-6082-2

The probability $P(A_m)$ that the fixed point will be covered by exactly m random objects is given by equation (4). For the expected values of the left and right hand side of this equation we have:

$$E[P(A_m)] = E\left[\sum_{i1,\ldots jm}\left(\psi_{i1}\psi_{i2}\ldots\psi_{im}\prod_{\substack{k=1\\k\neq i1;\ldots k\neq im}}^{n}(1-\psi_k)\right)\right] \tag{7}$$

Because the volume fractions ψ_k of the covering bodies are independent random variables, according to a well-known result in statistics, the expectation of a product of random variables is equal to the product of the expectations of the random variables. Consequently, equation(7) becomes:

$$\begin{aligned} &E[P(A_m)] = P(A_m) = \sum_{i1,\ldots jm} E\left(\psi_{i1}\psi_{i2}\ldots\psi_{im}\prod_{\substack{k=1\\k\neq i1;\ldots k\neq im}}^{n}(1-\psi_k)\right) = \\ &\sum_{i1,\ldots jm} E[\psi_{i1}]E[\psi_{i2}]\ldots E[\psi_{im}]\prod_{\substack{k=1\\k\neq i1;\ldots k\neq im}}^{n}(1-E[\psi_k]) = \binom{n}{m}\overline{\psi}^m(1-\overline{\psi})^{n-m} \end{aligned} \tag{8}$$

As a result, the expected fractions from the volume V, covered by the separate random objects with volumes v, are given by the separate terms of the binomial expansion of

$$[(1-\overline{\psi})+\overline{\psi}] = 1 \tag{9}$$

$$[(1-\overline{\psi})+\overline{\psi}]^n = \sum_{i=0}^{n}\binom{n}{i}\overline{\psi}^i(1-\overline{\psi})^{n-i} = 1 \tag{10}$$

The expected fraction of the volume covered by exactly m random objects is given by

$$P(A_m) = \binom{n}{m}\overline{\psi}^m(1-\overline{\psi})^{n-m} \tag{11}$$

The expected covered volume depends only on the mean volume of the covered objects and does not depend on the variance of the volumes of the covering objects or on their shape.

Now consider a case where n consumers demand a particular resource, during an operating period with length L. The durations of the demands from the consumers are d_i (i=1,...,n). The ratios of the durations of the demands from the separate consumers are given by $\psi_i = d_i / L$. The maximum number of consumers whose demand can be satisfied simultaneously by the sources is m.

Theorem 3. *If one source can satisfy only a demand from a single consumer at a time, the expected time fraction of unsatisfied demand related to m sources and n consumers is given by the expression*

 ISBN: 978-84-608-6082-2

$$p_{\geq m+1} = 1 - \sum_{i=0}^{m} \binom{n}{i} \overline{\psi}^{i} (1-\overline{\psi})^{n-i} \tag{12}$$

Proof. Unsatisfied demand related to *m* sources and *n* consumers (*n* > *m*) is present in the case where more than *m* consumers require a source simultaneously. Let p_0 denote the probability that a fixed point in the interval (0,*L)* will not sample any demand; p_1 denote the probability that the fixed point will sample exactly one random demand,...,and p_m denote the probability that the fixed point will sample exactly *m* random demands.

The probability $p_{\geq m+1}$ that the fixed point will sample more than *m* random demands, randomly placed in the time interval (0,*L*) is then given by

$$p_{\geq m+1} = 1 - (p_0 + p_1 + ... + p_m) \tag{13}$$

According to equation (10), the sum of the probabilities $p_0 + p_1 + ... + p_m$ is given by $\sum_{i=0}^{n} \binom{n}{i} \overline{\psi}^{i} (1-\overline{\psi})^{n-i}$. Hence, the theorem has been proved.■

Note that the sum in the right hand side of equation (12) is part of the binomial expansion of the expression $[(1-\overline{\psi})+\overline{\psi}]^n$, which is equal to unity (see equation 10). In equation (12), the term $(1-\overline{\psi})^n$ is the expected fraction of time during which no demand is present, the term $n(1-\overline{\psi})^{n-1}\overline{\psi}^1$ is the expected fraction of time during which exactly one random demand is present, the term $\frac{n(n-1)}{1\times 2}(1-\psi)^{n-2}\psi^2$ is the expected fraction of time during which exactly two random demands are present simultaneously,..., $\frac{n(n-1)...(n-m+1)}{1\times 2\times ...\times m}(1-\overline{\psi})^{n-m}\overline{\psi}^m$ is the expected fraction of time during which exactly *m* random demands are present simultaneously. The equations presented in this paper have been verified by a Monte Carlo simulation involving measuring and accumulating directly the multiple intersections. Because of lack of space, the details related to the algorithm have been omitted.

3. Determining the optimal number of consumers served by a specified number of sources

Consider a finite time interval during which a number of consumers demand a particular service independently and randomly, for a particular duration *d*. Constraint on the repair resources occurs if a random demand arrives while all sources are serving other consumers. Suppose that the maximum tolerable expected fraction of unsatisfied demand is α. Solving equation (14) with respect to *n* then yields the maximum number of consumers that can be serviced:

 ISBN: 978-84-608-6082-2

$$\alpha = 1 - \sum_{i=0}^{m} \binom{n}{i} \psi^i (1-\psi)^{n-i} \tag{14}$$

This equation can be solved by a repeated bisection.
Figure 2 gives the expected fraction of unsatisfied demand as a function of the number of consumers. The figure corresponds to a fraction $\psi = d/L = 0.15$ of an individual demand.

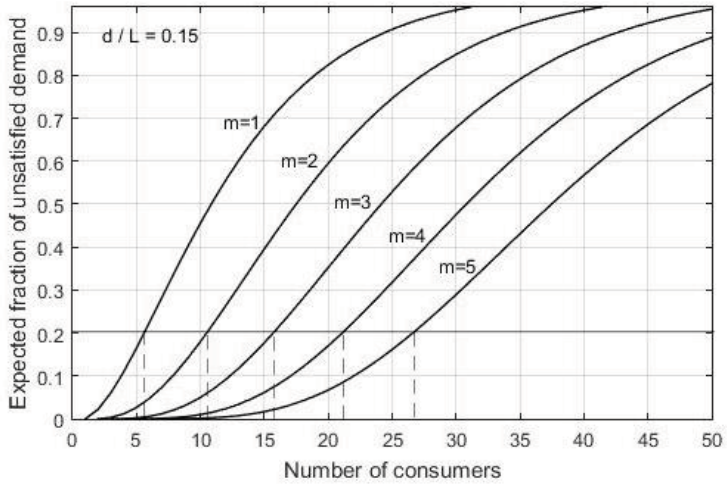


Figure 2. The expected fraction of unsatisfied demand as a function of the number of consumers, at different number of sources *m*.

By using repeated bisection, at a specified maximum acceptable level $\alpha = 10\%$ and $m = 3$ sources, the repeated bisection routine determined that at most 12 users can be serviced by the sources without exceeding the specified critical level $\alpha = 10\%$ of unsatisfied demand.

At a specified level of unsatisfied demand (for example $\alpha = 20\%$, the maximum number of consumers that can be satisfied can also be determined directly from the curve. As can be verified from the plots in Fig.2, increasing the number of sources sharply increases the number of consumers that can be serviced without exceeding the maximum tolerable fraction of unsatisfied demand.

4. Determining the optimal number of sources serving a given number of consumers

Suppose that the number of consumers *n* and the maximum tolerable expected fraction of unsatisfied demand α have been specified. Solving equation (14) with respect to *m* now yields the minimum number of sources required to service the *n* consumers such that the expected fraction of unsatisfied demand does not exceed α. Finding the minimum number of sources which guarantees a risk of unsatisfied demand below a specified level is critical in striking the right balance between risk

 ISBN: 978-84-608-6082-2

of unsatisfied demand and costs. If the sources are medical personnel, repairmen or simply extra equipment, increasing the number of sources unnecessarily means extra salary costs or investment, which undermines the profit. Too few sources means increased risk of unsatisfied demand, risk of fatalities, damage to health, dissatisfied customers, etc.

Again, equation (14) can be solved by a repeated bisection with respect to *m*, by keeping the number of customers n and the fraction $\psi = d/L$ constant. Alternatively, α versus *m* can be plotted and the smallest value *m* which still gives expected fraction of unsatisfied demand smaller than the specified value α can be selected.

Figure 3 gives the expected fraction of unsatisfied demand as a function of the number of sources for different ratios $\psi = d/L = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6$. The number of consumers is fixed ($n = 20$).

By using repeated bisection, at a specified maximum acceptable level $\alpha = 10\%$ and $n = 20$ consumers each characterised by a fraction $\psi = d/L = 0.2$, the repeated bisection routine determined that the minimum number of sources must be 6 for the expected fraction of unsatisfied demand to remain below 10%.

At a specified level of the expected fraction of unsatisfied demand (for example $\alpha = 20\%$, the minimum number of sources that can guarantee expected fraction of unsatisfied demand of $\alpha = 20\%$ or below can also be determined directly from the curve.

5. Expected fraction of unsatisfied demand for random demands following a homogeneous Poisson process on a time interval

Suppose that the random demands follow a homogeneous Poisson process with density λ, on a time interval with length *a*. The number of available sources is *m*. The durations of the random demands follow a particular distribution with mean μ. Unsatisfied demand is present only if the number of random demands is greater than the number of available sources *m*.

The expected fraction of unsatisfied demand given that there are exactly $n > m+1$ random demands on the time interval is given by

$$p_{\geq m+1} = 1 - \sum_{i=0}^{m} \binom{n}{i} \overline{\psi}^{i} (1-\overline{\psi})^{n-i} \tag{12}$$

For $n \leq m$, the expected fraction of random demand is zero.

For random demands following a Homogeneous Poisson process with density λ, on a time interval with length *a*, the probability that there will be exactly $m+1$ random demands on the time interval 0,*a* is given by $\frac{e^{-\lambda a}}{(m+1)!}(\lambda a)^{m+1}$.

The weighted fraction of unsatisfied demand given that there will be exactly $m+1$ random demands is therefore given by

 ISBN: 978-84-608-6082-2

$\frac{e^{-\lambda a}}{(m+1)!}(\lambda a)^{m+1}\left[1-\sum_{i=0}^{m}\binom{n}{i}\overline{\psi}^{i}(1-\overline{\psi})^{n-i}\right]$. The total weighted fraction of unsatisfied demand is therefore: $\sum_{n=m+1}^{\infty}\frac{e^{-\lambda a}}{(m+1)!}(\lambda a)^{m+1}\left[1-\sum_{i=0}^{m}\binom{n}{i}\overline{\psi}^{i}(1-\overline{\psi})^{n-i}\right]$.

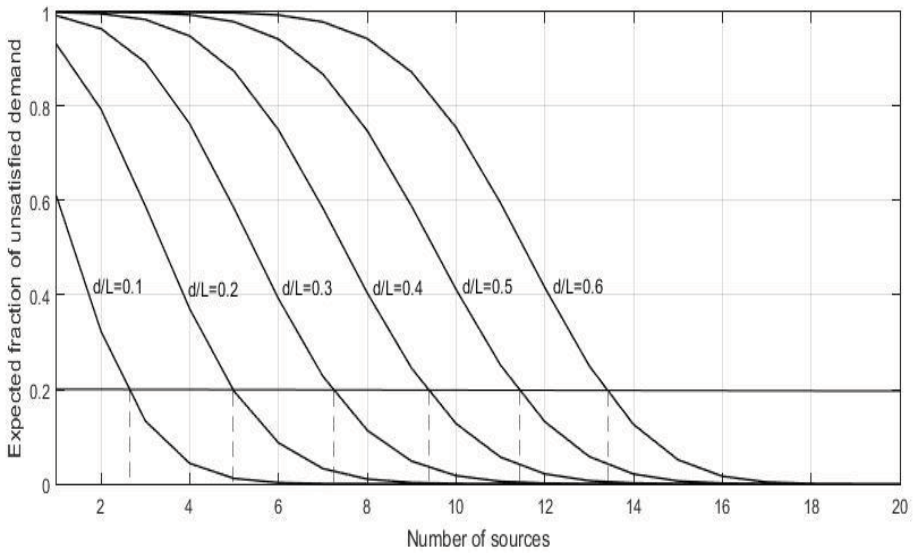


Figure 3. The expected fraction of unsatisfied demand as a function of the number of sources, at different ratios $\psi = d / L$.

Consider a complex system (production system, computer network, etc.) where the random demands are the failures of the separate components building the system. The presented computational framework can then be applied to determine the maximum size of the system that can be serviced by a given number of repair people and also the minimum number of repair people needed to service a system of given size.

6. Conclusions

• A computational framework has been created for evaluating the risk of unsatisfied demand arising from random demands satisfied by multiple sources.
• At the heart of the proposed computational framework is a general equation evaluating the risk of unsatisfied demand by the expected fraction of time of unsatisfied demand. The equation covers the important case of multiple sources servicing multiple random demands.
• Based on the closed-form solution, an efficient optimisation method has been developed for determining the maximum number of consumers that can be serviced by a given number of sources, such that the risk of unsatisfied demand remains below a maximal tolerable level.

 ISBN: 978-84-608-6082-2

• Based on the closed-form solution, an efficient optimisation method has been developed for determining the minimal number of sources needed to service a specified number of consumers such that the risk of unsatisfied demand remains below a maximal tolerable level.

7. References

[1] Allen A.O., *Probability, Statistics and Queuing theory with computer science applications*, 2nd ed., Academic press Inc., 1990.

[2] Calin O., K.Udriste, *Geometric modelling in probability and statistics*, Springer, 2014.

[3] Coffman E.G., Jr.,1 L. Flatto,1 P. Jelenkovic, and B. Poonen, *Packing Random Intervals On-Line*, *Algorithmica* **22** (1998) 448–476.

[4] Coffman Jr E.G, C.L. Mallows and B.Poonen, *Parking arcs on the circle with applications to one-dimensional communication networks*, The Annals of Applied Probability, **4**(4) (1994) 1098-1111.

[5] Genedenko B.V., I.N.Kovalenko, *Introduction to queuing theory*, 2nd ed., Birkhauser, Berlin, 1989.

[6] Giambene G., *Queuing theory and telecommunications*, Networks and applications, Springer, 2005.

[7] Gross, D., & Harris, C. M. *Fundamentals of queuing theory* (2nd ed.). New York: Wiley, 1985.

[8] Huffer F.W. and L. A. Shepp, *On the Probability of Covering the Circle by Random Arcs*, Journal of Applied Probability, **24** (2) (1987) 422-429.

[9] Justicz J., E. R. Scheinerman, and P. M.Winkler. *Random intervals*. Amer. Math. Monthly, **97** (1990) 881-889.

[10] Kalashnikov V., *Mathematical methods in queuing theory*, Springer-Science+Business Media, 1994.

[11] Khintchine, A. Y. *Mathematical methods in the theory of Queueing*. London: Griffin., 1969.

[12] Shepp L.A., *Covering the circle with random arcs*, Israel J.Math., **11** (1972) 328-345.

[13] Solomon H., *Geometric probability*, Society for industrial and applied mathematics, 1978.

[14] Stevens W.L., Solution to a geometrical problem in probability, *Ann.Eugenics*, **9** (1939) 315-320.

[15] Stroock D.W., *Probability theory*: An analytical view, 2nd ed., Cambridge University press, 2011.

[16] Thompson W.A., *Point process models with applications to safety and reliability*, Chapman & Hall, 1988.

[17] Venkatesh S.S., *The theory of probability: explorations and applications*, Cambridge University Press, 2013.

[18] Weiss N.A., *A course in probability*, Pearson/Addison Wesley, 2006.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Numerical Solutions of GPE under Gaussian Trap

Eren Tosyali[1] and Fatma Aydogmus [2]

[1] *Vocational School of Health Services, Istanbul Bilgi University, Istanbul, TURKEY*
[2] *Department of Physics, Istanbul University, Istanbul, TURKEY*

emails: eren.tosyali@bilgi.edu.tr
fatma.aydogmus@gmail.com, fatmaa@istanbul.edu.tr

Abstract

In this study, we investigate the regular and chaotic solutions of a BEC system in 1D tilted Gaussian optical lattice potential by constructing their Poincaré sections in phase space depending on the system parameters.

Key words: Bose-Einstein Condensate, Gross-Pitaevskii Equation, Optical Lattice Potential, Chaos.
MSC2000: AMS Codes (optional)

1. Introduction

The condensate is well described by a mean field theory and a macroscopic wave function , solving the nonlinear Schrödinger equation so-called Gross– Pitaevskii equation (GPE) [1, 2, 3] that includes a nonlinear term representing particle-particle interactions [4, 5]. GPE is an integrable system and the integrability could be easily broken by external potentials of different forms [4, 5]. So it is difficult to obtain its exact solutions with external potential analytically. Therefore we perform numerical simulations in this paper. We investigate the regular and chaotic solutions of the BEC system in the tilted Gaussian optical lattice potential. As is well known, the main traditional way for checking chaos is the construction of Poincaré section [6, 7, 8, 9, 10, 11] which provides regular and chaotic behaviour regions. Therefore we construct the Poincaré sections in phase space depending system parameters.

 ISBN: 978-84-608-6082-2

2. Model

In the low temperature regime, a BEC is well described by the non- linear Schrödinger equation known as the Gross–Pitaevskii Equation (GPE) with the macroscopic wave function $\psi = \psi(x,t)$ which evaluates with time and space [12, 13]. The 1-D Gross-Pitaevskii equation is given as

$$i\hbar\psi(x,t) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t) + \left[V_{ext}(x) + g_0\left|\psi(x,t)\right|^2\right]\psi(x,t) \tag{1}$$

where m is the mass of the atoms of the condensate, g_0 describes the interaction between atoms in the condensate and given by $g_0 = 4\pi\hbar^2 a_s/m$ where a_s is s-wave scattering length between atoms. It is positive for repulsive interaction and negative for attractive interaction (in our case $a < 0$). $V_{ext}(x)$ is the external trapping potential. We choose the external trapping potential as below,

$$V_{ext}(x) = V(x) + Fx \tag{2}$$

where $V(x)$ is the optical potential and F the inertial force. This force, which generates a tilted optical potential, accelerates the atoms in the x direction and leads to the atoms tunnelling out of the traps [4, 14]. In this paper, we investigated 1-D Gross-Pitaevskii equation under the tilted Gaussian optical lattice potential. We construct a potential which involve Gaussian peaks along the x direction in spatial phase by using Fourier transform procedure. Each Gaussian peaks described by Eq.3. A is amplitude of each Gaussian peaks and μ and σ are the system parameters.

$$f(x) = Ae^{-\frac{(x-\varepsilon)^2}{2\sigma^2}} \tag{3}$$

In order to generate Gaussian pulse potential, we define a step length

$$xx = \frac{x_{\max} - x_{\min}}{n} \tag{4}$$

 ISBN: 978-84-608-6082-2

$x_{\max}$ and $x_{\min}$ are maximum and minimum border for numerical calculation. n defines the step length. We create B matrix by evaluating $f(x)$ from $x_{\max}$ to $x_{\min}$ as below.

$$B = \begin{pmatrix} x_{\min} & Ae^{-\frac{(x_{\min}-\mu)^2}{2\sigma^2}} \\ \vdots & \vdots \\ x_{\max}-tt & Ae^{-\frac{(x_{\max}-tt-\mu)^2}{2\sigma^2}} \end{pmatrix} \tag{5}$$

The discreet Fourier Transform of B matrix is given in Eq. 6

$$C = \begin{pmatrix} Ae^{-\frac{(x_{\min}-\mu)^2}{2\sigma^2}} e^{-2\pi i x_{\min} k/n} \\ \vdots \\ Ae^{-\frac{(x_{\max}-xx-\mu)^2}{2\sigma^2}} e^{-2\pi i (x_{\max}-tt) k/n} \end{pmatrix} \tag{6}$$

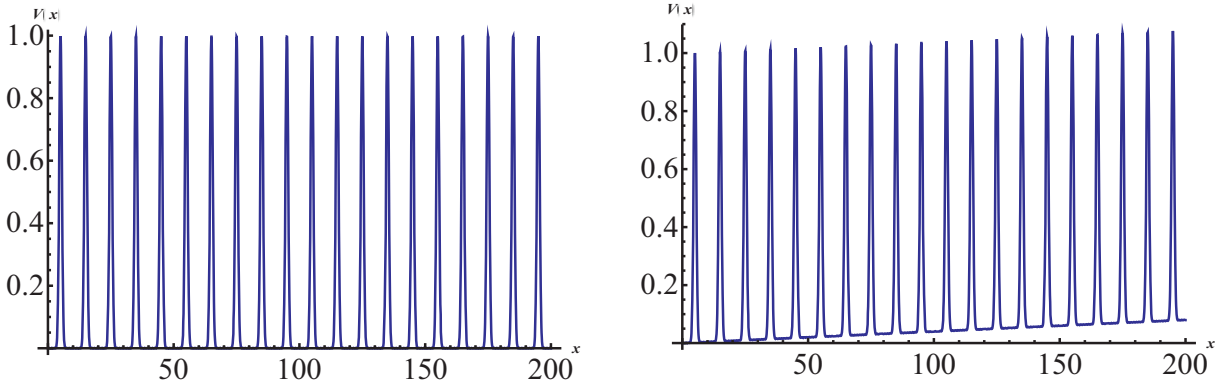
From C matrix, we find Eq.7 that generates one dimensional Gaussian pulse potential.

$$V(x) = \frac{C_1}{\sqrt{n}} + \frac{2}{\sqrt{n}} \sum_{j=2}^{j_{\max}} Abs\left(C_j\right) \mathrm{Cos}\left(w \frac{j-1}{xx} x - Arg\left(C_j\right) \right) \tag{7}$$

here $j_{\max} = \frac{3tt}{2\pi\sigma} + 1$

In Fig. 1 we show Gaussian pulse potential a) without tilted term, b) with small tilted term.

 ISBN: 978-84-608-6082-2



(a) (b)

Figure 1: Plot the Gaussian pulse potential with the parameters $A=1$, $n=90$, $\mu=50$, $\sigma=5.26$, $w=0.02$, $t_{\min}=0$, $t_{\max}=200$, $j_{\max}=10.071$ (a) $F=0$ and (b) $F=0.0004$

In order to obtain a simple description and a better understanding of the BEC dynamics, we consider Ψ as [12]

$$\Psi(x,t)=\Phi(x)e^{\frac{i\mu t}{\hbar}} \tag{8}$$

Here μ is the chemical potential of the condensate and $\Phi(x)$ is a real function independent of time. $\Phi(x)$ is normalized to the total number of particles in the system, i.e.,

$$\int|\Phi(x)|^2\,dx=N \tag{9}$$

where N is the particle number.

Substitution of Eqs.(3) and (4) into Eq.(1)

$$\mu\Phi(x)=-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Phi(x)+\left[V(x)+Fx+g_0\Phi|(x)|^2\right]\Phi(x) \tag{10}$$

which can also be written in the following form

$$\frac{d\Phi}{dx} = \left[V(x) + \xi x - \gamma + \eta \left| \Phi \right|^2 \right] \Phi \tag{11}$$

Where $\left(\nu(x) = \frac{2mV(x)}{h^2}, \gamma = \frac{2m\mu}{\hbar^2}, \xi = \frac{2mF}{\hbar^2}, \eta = \frac{2mg_0}{\hbar^2} \right)$. Setting the solution of Eq. (7) of form

$$\Phi(x) = \phi(x) e^{i\theta(x)} \tag{12}$$

Inserting Eq. (8) into Eq. (7),

$$\frac{d^2\phi}{dx^2} = \phi \left(\frac{d\theta}{dx} \right)^2 + \left[\nu(x) + \xi x - \gamma + \eta \left| \phi \right|^2 \right] \phi \tag{13a}$$

$$\frac{d}{dx} \left(2\phi^2 \frac{d\theta}{dx} \right) = 0 \tag{13b}$$

Eq. (13b) denotes the existence of a flow density,

$$J = 2\phi^2 \frac{d\theta}{dx} \tag{14}$$

If we put J into Eq. (13a), we have a nonlinear equation as below.

$$\frac{d^2\phi}{dx^2} = \frac{J^2}{4\phi^3} + \left[\nu(x) + \xi x - \gamma + \eta \left| \phi \right|^2 \right] \phi \tag{15}$$

It is difficult to obtain the exact solution of Eq. (11) due to its complexity therefore numerical solutions were performed. For numerical solution we can reduce Eq. 15 in to first order coupled equations.

$$\dot{\phi}_1 = \phi_2 \tag{16a}$$

$$\dot{\phi}_2 = \frac{J^2}{4\phi_1^3} + [\frac{C_1}{\sqrt{n}} + \frac{2}{\sqrt{n}} \sum_{j=2}^{j_{\max}} Abs\left(C_j\right) \mathrm{Cos}\left(w \frac{j-1}{xx} x - Arg\left(C_j\right)\right) + \tag{16b}$$

$$\xi\phi_3 - \gamma + \eta\left|\phi_1^2\right|]\phi_1$$

$$\dot{\phi}_3 = \Omega \tag{16c}$$

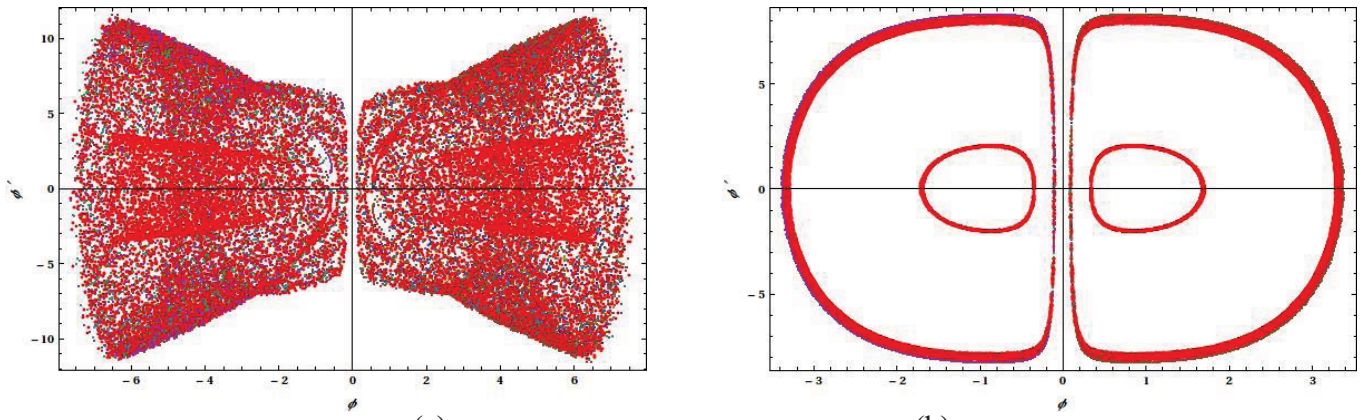
If we take $w = \xi = \eta = J = 0$ for Eq.16(b), we find

$$\dot{\phi}_1 = \phi_2 \tag{17a}$$

$$\dot{\phi}_2 = [\frac{C_1}{\sqrt{n}} + \frac{2}{\sqrt{n}} \sum_{j=2}^{j_{\max}} Abs\left(C_j\right) \mathrm{Cos}\left(-Arg\left(C_j\right)\right) - \gamma - 1\left|\phi_1^2\right|]\phi_1 \tag{17b}$$

3. Regular and Chaotic Numerical Solutions

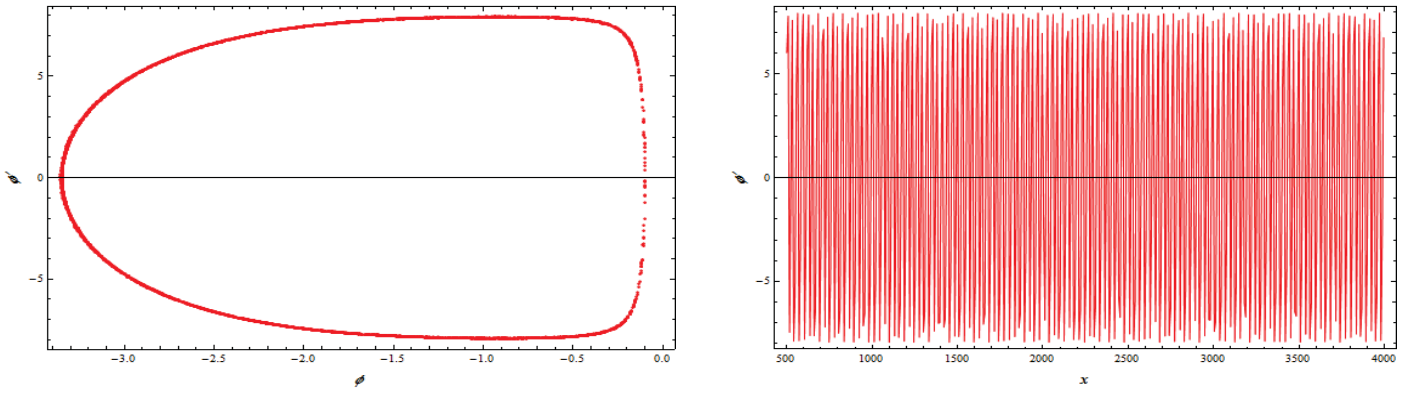
In this section, we investigate the regular and chaotic solutions of the BEC system with attractive interaction in different parameter regions. We consider the Gaussian trap that the system exhibits regular and chaotic behaviours under it. Fig.2 shows that the Poincaré sections with twenty different initial conditions for the parameters set $J = 0.2$, $\upsilon_1 = 1$, $\upsilon_2 = 0.8$, $w_1 = 2\pi$, $w_2 = 5\pi$, $\gamma = 0.5$ initial condition: $\phi[0] = -0.1$, $\phi'[0] = 0.9$. Fig.2(a) display the regular state distribution on the complete phase space for $\xi = 0.01$. As we weaken flow density, Fig.2(b) exhibits chaotic orbits.



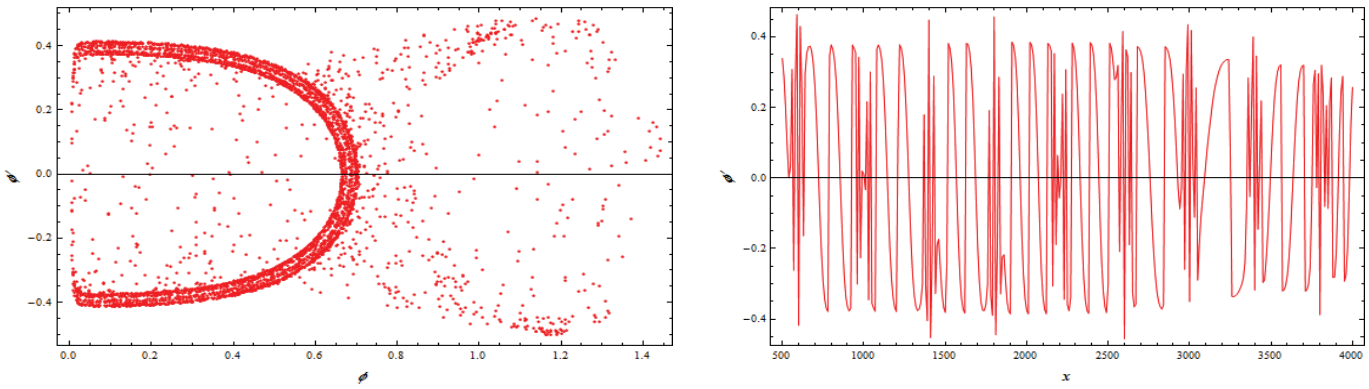
(a) (b)

Figure 2: (a) Regular phase space of BEC with the Gaussian optical lattice potential for $\xi = 0.01$ (b) Chaotic phase space for $\xi = 0.000001$ Parameters set: $A = 0.2, J = 0.8, \gamma = 0.5$, $\eta = -1$ with twenty different initial condition $n = 90$, $\mu = 50$, $\sigma = 5.26$, $t_{\min} = 0$, $t_{\max} = 4000$

In Fig.3 we see the phase space displays and time series plots of system under the Gaussian optical lattice potential for the one initial condition $\phi[0] = 0.1$, $\phi'[0] = 0.4$ and $\xi = 0.00001, \gamma = 0.5$. The system exhibits regular behaviours but when the flow density is taken very small the phase orbit on the Poincaré section become aperiodic and chaotic, respectively, as shown in Figs. 3(a) and 3(b).



(a)



(b)

Figure 3: (a) Regular phase space and time series plot of BEC with the Gaussian optical lattice potential for $A = 0.1$, $J = 0.8$. (b) Chaotic phase space and time series plot for $A = 1.5$, $J = 0.002$. Parameters set: $J = 0.2$, $\xi = 0.00001$, $\gamma = 0.5$ and $\phi[0] = 0.1$, $\phi'[0] = 0.4$, $n = 90$, $\mu = 50$, $\sigma = 5.26$, $t_{\min} = 0$, $t_{\max} = 4000$.

4. Conclusion

In summary, we have studied the BEC with the tilted Gaussian optical lattice potential. The regular and chaotic numerical solutions of system are investigated by constructing the Poincaré sections depending on flow density and tilted term. Numerical solutions indicate that there exists chaos in the system. These instabilities of the system refer to negative interatomic interactions.

5. References

[1] S.N. Bose, Z. Phys. 26 (1924) 168..
[2] A. Einstein, Sitzungsber. K. Preuss. Akad. Wiss. 261 (1925) 3.
[3] M.H. Anderson, J.R. Ensher, M.R. Matthews, C.E. Wiemann, E.A. Cornell, Science 269 (1995) 198.
[4] Q. Thommen, J. Claude and V. Zehnle, Phys. Rev. Lett. 91 (2003) 210405.
[5] D. Zhao, H.G. Luo and H.Y. Chai, Phys. Lett. A 372 (2008) 5644.
[6] B. Hu and L. M. Kuang Phys. Rev. A 61 (2000) 023604.

 ISBN: 978-84-608-6082-2

[7] Q. T. Xie, W. H. Hai and G.S. Chong, Chaos (2003) 13801.
[8] Cohen-Tannoudji CN, in: Les Prix Nobel 1997 (The Nobel Foundation, Stockholm,1998), pp. 87–108, Reprinted in: Rev. Mod. Phys. 70:707–719.
[9] K. B. Davis, M. O. Mewes, M.R. Adrews, N. J. Van Druten, D.S. Durfee, D. M. Kurn, W. Ketterle, Phys. Rev. Lett. 75 (1995) 3969.
[10] D.S. Jin, J.R. Ensher, M.R. Matthews, C.E. Wiemann, E.A. Cornell, Phys. Rev. Lett. 77 (1996) 420.
[11] C.C. Bradley, C.A. Sackett, J.J. Tollett, G. Hulet, Phys. Rev. Lett. 75 (1995) 1687.
[12] L.P. Pitaevskii, S. Stringari, Clarendon Press, Oxford, New York, ISBN: 0198507194, (2003).
[13] E.P. Gross, Nuovo Cimento 20 (1961) 454.
[14] F. Jıan-Shu, L. Xiang-Ping, Chin. Phys. B 20 (2011) 040310.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

A mathematical ranking model in learning analytics

A. van der Merwe[1], H.A. Kruger[1] and J.V. du Toit[1]
[1] School of Computer-, Statistical-, and Mathematical Sciences, North-West University, Potchefstroom Campus, South Africa

emails: Annette.VanDerMerwe@nwu.ac.za, Hennie.Kruger@nwu.ac.za, Tiny.DuToit@nwu.ac.za

Abstract

The introduction of new educational trends triggers the development of innovative methods to collect, analyse and report the data subsequently generated. This paper discusses the implementation of a multi-stage mathematical class ranking model as a method in learning analytics, applied to a Computer Science module. The model applies the principle of Pareto optimality in an outputs-only data envelopment model to categorise students into classes of similar efficiency which are ranked according to dominance. The model is then adapted to calculate improvement targets for each student.

Key words: class ranking, data envelopment analysis, learning analytics, Pareto optimality, student ranking.

1. Introduction

Development in educational technology has seen traditional learning models in instruction gradually being replaced with modern trends like blended- and online learning [1]. These trends include the use of learning management systems, sites on social media networks and the incorporation of learning analytics (LA). LA includes the real-time collection of data, or "big data", which are so large that conventional collection methods have proved inadequate [2]. It also encompasses the use of analyses on big data to make multi-criteria decisions, report and inform in the educational system [3]. While blended learning methods in education are gaining

 ISBN: 978-84-608-6082-2

popularity and being implemented more frequently, the process of accurately measuring student progress is proving to be somewhat of a challenge.

LA presents an important instructional application in the form of monitoring individual student performance [4]. The timely collection of data also enables lecturers to measure the progress of students at any stage of instruction. It has been found that regularly informing students on their academic progress and ranking in relation to their peers, stimulates a competitive attitude which effectively motivates them to try and outperform their class mates [5]. LA also enables lecturers to use predictive modelling techniques which can:

- Provide students with early academic outcome alerts [4]; and
- Help identify and provide interventions for academically at-risk students at an early stage [3].

In this paper, a multi-stage class ranking model will be implemented as a LA method to rank students and provide improvement intervention targets. The model uses a data envelopment analysis (DEA) approach which is a special linear programming application in which the performance efficiency of a specified number of operating units with the same objectives is measured relative to one another [6].

In the next section, the concept of Pareto optimality and how it relates to this work is discussed and the mathematical model used for DEA is formulated and explained. This is followed by an outline of the dual-formulation used to determine intermediate improvement targets for students. Section 3 describes the application of the multi-stage mathematical model on a data set of a Computer Science module, followed by a discussion of the results and some concluding remarks.

2. Pareto optimality and DEA

DEA is commonly implemented in cases where different operating units, or decision making units (DMUs), need to be compared with one another. Typical examples include banks, hospitals, fast-food outlets within a certain chain, or in this case, students according to academic performance.

DEA compares a certain number of DMUs based on their inputs or resources and their outputs [7], by calculating an efficiency score that represents the relationship between its inputs and outputs. The result gives an indication of whether a particular unit is less efficient than the others. It is however very difficult to determine fair, comparable inputs

where students are concerned. Consider for instance the accuracy of data relating to the number of hours study time per student, or lifestyle factors.

A similar problem was addressed by Kao and Lin [8] who implemented a DEA model without inputs to categorise Management Colleges into classes of differing levels. All colleges in the same class were considered equally efficient based on their outputs, and could not be compared to one another. This class ranking model will be applied to arrange students into different classes of efficiency, where all students in the same class are considered equally efficient. As far as can be ascertained, the DEA outputs-only class ranking has not previously been used on students in such an application.

One of the advantages of this model is that the principle of Pareto optimality applies, meaning that Pareto-efficient students are categorised into one class, and they dominate those in lower classes. In a traditional DEA output-oriented method, a DMU, or in this case a student, is considered Pareto-efficient if none of the output scores can increase without decreasing another one of the output scores or without increasing at least one of the input scores [9]. In mathematical terms, student k_0 is Pareto-efficient if no other feasible student $k \neq k_0$ exists so that $y_{kj'} > y_{k_0 j'}$ for some output criteria j' and $y_{kj} > y_{k_0 j}$ where $j \neq j'$ and $x_{kr} \leq x_{k_0 r}$ for all input criteria r, where y_{kj} are the output levels for student k, with $j = 1, \ldots, s$, x_{kr} are the input levels for student k, and $r = 1, \ldots, m$.

For this application, it means that all students considered to be Pareto optimal should be categorised as equally efficient and receive the same ranking, while those they dominate must be ranked lower. Consider for example, a data set of five students, denoted by A to E. These students must be evaluated on two outputs, which are scores for class attendance and class assignments, shown in Table 1.

Table 1: Example data set

	Output1 (Attendance)	Output2 (Assignments)	Total
Student A	20	70	90
Student B	80	70	150
Student C	65	60	125
Student D	50	90	140
Student E	50	50	100

 ISBN: 978-84-608-6082-2

If the two criteria are considered equally important, it would mean that student B is the most efficient, followed by students D, C, E, and A respectively. However, student performance is rarely measured with criteria that are considered equally important. Attendance cannot be compared to assignments, so it will be more rational to search for those students who dominate others. This is done by calculating an efficiency score for each student. The dominance relationship of the students represented by the data set, are shown in Figure 1.

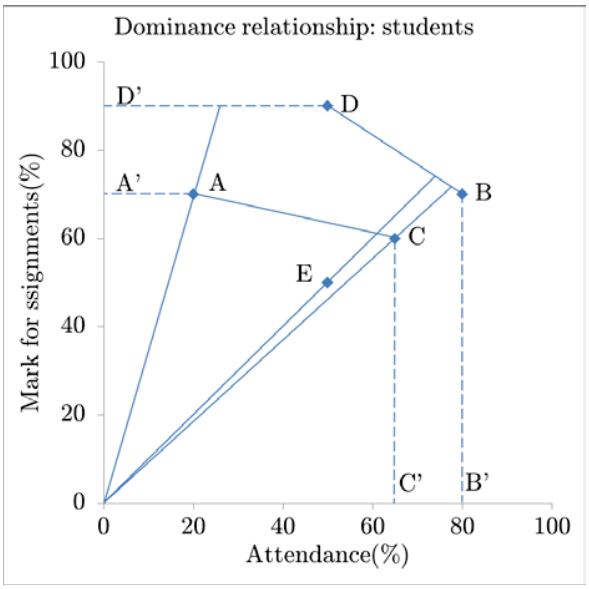


Figure 1: Dominance relationship of the example data set

Students D and B lie on the Pareto optimal frontier, denoted by line segments D'DBB', and are called Pareto optimal because they dominate students A, C, and E. The dominant students are categorised into the highest class and have the highest ranking. Students A and C are dominated because they lie inside the frontier and are categorised into a lower class. Likewise, student E is dominated by students A and C and is in the lowest class. This means that students D and B who are members of the dominant set will have the highest class ranking, followed by students A and C, and student E will have the lowest class ranking. Although student E had a higher total score than student A, he/she was categorised into a lower class. This happens when a student performs predominantly well in all the criteria then the equal weights method results in a higher ranking, whereas the dominance method favours a student who performs exceptionally well in a specific criterion.

Finding the dominant sets for calculating the efficiency score, or composite index E_k, is done by implementing a basic DEA model without inputs. Pareto optimal students are therefore determined using the following simple DEA output-only model [8]:

$$\text{Maximise} \quad E_k = \sum_{j=1}^{m} y_{kj} w_j \qquad (1)$$

$$\text{subject to} \quad \sum_{j=1}^{m} y_{ij} w_j \leq 1, \qquad i = 1, \ldots, n \qquad (2)$$

$$w_j \geq \varepsilon > 0, \qquad j = 1, \ldots, m \qquad (3)$$

where E_k is the composite index for every student k, y_{ij} is the output measure of student i in criterion j, w_j is the weight for criterion j, n is the number of students to be evaluated, m is the number of criteria, and ε is a small positive number.

The purpose of constraint (3) is to ensure that all outputs are considered when calculating student efficiency, which means that no weights are allowed to be zero. For each student, the optimal weights are selected for calculating the composite index E_k. If $E_k = 1$, student k is considered Pareto optimal, categorised as a member of the current dominating set and excluded from further calculations. If $E_k < 1$, then student k is Pareto non-optimal and remains in the list of students still to be categorised. This process is repeated k times, once for each of the students. Table 2 compares the rankings of the equal weights method to the classes in the dominance method.

Table 2: Equal weight rankings vs. dominance class rankings

Equal weight method		Dominance method	
Ranking	Student	Class	Student
1	B	1	B
2	D	1	D
3	C	2	A
4	E	2	C
5	A	3	E

Here the difference in ranking is evident in the case of student A, who performed poorly in output 1 and good in output 2. The equal weights method ranked student A 5th, whereas the dominance method ranked this student into class 2.

The algorithm for determining Pareto optimal students and the subsequent class ranking [10], according to model (1)-(3) follows.

Algorithm 1: Class ranking multiple DMUs

Input: Let $N = \{DMU_1, \dots, DMU_n\}$ be the set of n DMUs, and $\alpha \in \mathbb{Z}^+$.
Output: A number of sets, $C_1, \dots, C_f$, where all units in C_i have the same ranking, and are ranked higher than those in C_j for $i, j \in \{1, \dots, f\}$ and $i < j$.

```
1   f ← 1;
2   while N ≠ ∅ do
3   |  Initialise C_f
4   |  α ← 1;
5   |  while α ≤ n do
6   |  |  k ← DMU_α;
7   |  |  Apply model (1)-(3) and calculate E_k
8   |  |  if E_k = 1 then
9   |  |  |  Add unit k to C_f
10  |  |  |  Delete unit k from N
11  |  |  end
12  |  |  α ← α + 1;
13  |  end
14  |  f ← f + 1;
15  end
```

Algorithm 1 will repeat until all the students have been categorised into classes. Although this stage determines the classes and resulting class rankings, students can only improve their positions if they knew which criteria need to be improved and by how much. To provide them with such intervention actions, intermediate targets must be set for each student. This is done using the dual formulation of model (1)-(3) [8]:

$$\text{Minimise} \quad E_k = \sum_{i=1}^{n} \lambda_i - \varepsilon \sum_{j=1}^{m} s_j \tag{4}$$

$$\text{subject to} \quad \sum_{i=1}^{n} y_{ij}\lambda_i - s_j = y_{kj}, \qquad j = 1, \dots, m \tag{5}$$

$$\lambda_{ji}, s_j \geq 0, \qquad \begin{matrix} i = 1, \dots, n; \\ j = 1, \dots, m \end{matrix} \tag{6}$$

where E_k is the composite index, y_{ij} is the output measure of student i in criterion j, n is the number of students to be evaluated, m is the number

of criteria, and ε is a small positive number. Denote $\theta = \sum_{i=1}^{n} \lambda_i$, then $(y_{kj} + s_j)/\theta$ is the target, and $j = 1, \ldots, m$.

Although the calculated efficiency E_k is the same as in model (1)-(3), model (4)-(6) presents student k with targets to make an improvement, providing that the student is not a member of the highest class already. If a student belongs to the f^{th} class with $f \neq 1$, this model calculates $f - 1$ targets to show how much the student must improve in each criterion to be categorised in the next higher class.

The multi-stage model is therefore implemented in two phases: calculating the Pareto optimal students and categorising them into classes using model (1)-(3), and determining intermediate targets for all dominated students to improve their class rankings using model (4)-(6). The implementation of this multi-stage model on a data set of marks for students in a Computer Science module follows.

3. Implementation of the multi-stage mathematical model

The multi-stage model was implemented to obtain the class ranking of students according to their academic progress in a Computer Science module. The data set contains four criteria for 26 students enrolled in the module. The criteria are average scores for assignments, attendance, formative assessments, and summative assessments. Table 3 presents the four initial output criteria for each student, the total of the four criteria, and the class ranking resulting from Algorithm 1.

Table 3: Average output scores (rounded), total of output scores, and resulting class ranking of the students

Student	Output1 Assignments	Output2 Attendance	Output3 Formative assessments	Output4 Summative assessments	Total	Class
2	98	100	100	95	393	1
3	75	100	97	83	355	2
9	88	100	93	92	373	2
10	74	100	93	68	335	3
13	66	89	77	87	319	3
23	62	100	85	72	319	3
24	77	100	90	65	332	3
1	14	63	73	65	215	4
6	58	89	75	63	285	4
11	65	100	88	53	306	4

 ISBN: 978-84-608-6082-2

12	64	100	73	57	294	4
17	68	78	90	58	294	4
19	41	100	73	63	277	4
8	54	88	58	44	244	5
15	24	89	83	57	253	5
18	43	100	63	58	264	5
20	52	89	60	55	256	5
21	36	75	85	59	255	5
25	26	67	83	60	236	5
26	25	100	65	55	245	5
5	41	78	62	57	238	6
7	35	100	58	22	215	6
14	53	78	10	44	185	6
22	43	89	40	54	226	6
16	26	78	52	46	202	7
4	21	78	35	35	169	8

The students are categorised into 8 classes, which are ranked in order of dominance. The number of students in each class is shown in Figure 2.

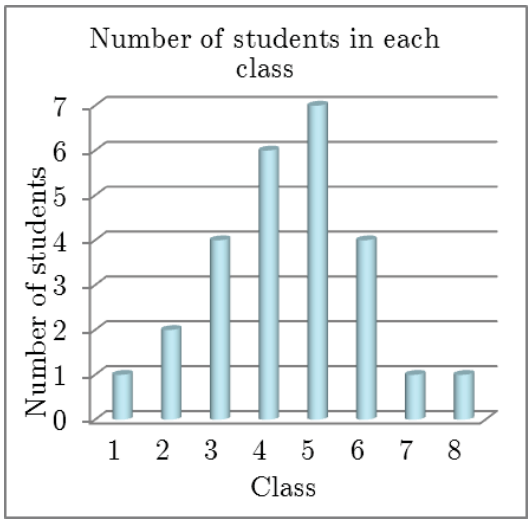


Figure 2: Number of students categorised into each class

From Table 3 it follows that student 2 is categorised into class 1 and therefore ranked higher than all the others. Students 3 and 9 are members of class 2, which means they are dominated by student 2, and themselves dominate the rest of the students. The students in classes 1 to 3 performed

well in all the criteria but especially so in output 2, attendance. The class ranking of these groups will therefore not differ very much from an equal-weighted ranking. However, a student with a higher total score can be dominated by one with a lower total score. Consider for example the data of students 1 and 5 which are repeated in Table 4.

Table 4: Comparison between students 1 and 5

Student	Output1 Assignments	Output2 Attendance	Output3 Formative assessment	Output4 Summative assessment	Total	Class
1	14	63	73	65	215	4
5	41	78	62	57	238	6

Student 5 is a member of class 6 and has a total score of 238, which is much higher than that of student 1 who is a member of class 4 and has a total score of 215. The dominance method here favours student 1 who has performed well in outputs 3 and 4, and poorly in output 1, as opposed to student 5 who had rather average scores for all those outputs.

The second stage of the model provided each student with intermediate targets which, if reached should categorise the student into a higher class. The targets are in the form of scores for each of the criteria. Suppose for example, that student 5 who is a member of class 6 wants to improve his/her class level, then targets are calculated by solving model (4)-(6). These targets for student 5 are shown in Table 5.

Table 5: Intermediate targets for class level improvement for student 5

Class level	Output1 Assignments	Output2 Attendance	Output3 Formative assessment	Output4 Summative assessment
6 (current)	41	78	62	57
5	42	96	66	58
4	45	87	74	63
3	62	100	85	73
2	57	100	79	73
1	67	100	79	73

The current class of student 5, with output scores of 41, 78, 62, and 57 respectively, is level 6. If the student wants to advance to level 5, output 1

needs to improve from 41 to 42, output 2 needs to improve from 78 to 96, and so forth.

This LA technique presented students with early academic outcome alerts, helped to identify academically at-risk students and also provided real-time intervention targets which students could use for improvement.

4. Conclusions

The increased use of blended- and online learning techniques in education, leads to learning management systems and LA progressively being implemented to manage and supervise data resulting from such modules. One of the requirements of LA is that lecturers must provide students with real-time feedback on their progress in their studies. This led to the search for novel and improved methods for measuring student progress and performance while also calculating reachable targets for student improvement.

This paper discussed the successful implementation of a multi-stage mathematical ranking model in a Computer Science module. The first stage of the model used the principle of Pareto optimality implemented in an output-only DEA model to categorise the students into classes of similar efficiency which were then ranked according to dominance. The second stage implemented the dual-formulation of the DEA model which provided each student with an intermediate target per relevant criterion, for class level improvement.

The mathematical ranking model implemented in this study enabled the comparison of students in terms of incomparable criteria. Students who had the same totals were for instance not necessarily categorised into the same class and students in the same class could not be compared. Although they could see which criteria they needed to improve for increased efficiency, not all of the targets could in reality be reached. Consider for example a formative assessment which generally constitutes a semester test. Depending on the lecturer, the mark cannot change after the test has been taken.

The addition of features which would allow students within the same class to be compared, as well as determining intermediate targets for only some of the criteria while others remain fixed, are currently being investigated.

5. References

[1] J. A. REYES, *The skinny on big data in education: Learning analytics simplified*, TechTrends: Linking Research & Practice to Improve

Learning, vol. 59, no. 2, pp. 75-79, March-April 2015.

[2] K. Sin and L. Muthu, *Application of big data in education data mining and learning analytics - A literature review*, ICTACT Journal on Soft Computing: Special Issue on Soft Computing Models for Big Data, vol. 5, no. 4, pp. 1035-1049, July 2015.

[3] J. Fiaidhi, *The next step for learning analytics*, IT Professional, pp. 4-8, September-October 2014.

[4] A. G. Picciano, *Big data and learning analytics in blended learning environments: Benefits and concerns*, International Journal of Artificial Intelligence and Interactive Multimedia: Special Issue on Multisensor User Tracking and Analytics to Improve Education and other Application Fields, vol. 2, no. 7, pp. 35-43, 2014.

[5] J. V. Du Toit, *Using participation marks to manage, motivate and inform on academic progress*, in: Conference for Outstanding Teaching/Learning and Innovative Technology Use, Potchefstroom, South Africa, 2015.

[6] D. R. Anderson, D. J. Sweeney, T. A. Williams, J. D. Camm, J. J. Cochran, M. J. Fry, and J. W. Ohlmann, *An introduction to management science: Quantitative approaches to decision making*, 14th ed. Boston, USA: Cengage Learning, 2014.

[7] B. W. Taylor III, *Introduction to Management Science*, 11th ed. Harlow: Person Education Limited, 2013.

[8] C. Kao and P-H. Lin, *Class ranking of the managament colleges in Taiwan*, Lecture Notes in Management Science, vol. 1, pp. 129-140, 2008.

[9] E. Thanassoulis, Introduction to the theory and application of data envelopment analysis: A foundation text with integrated software. Norwell, USA: Kluwer Academic Publishers, 2001.

[10] L. M. Seiford and J. Zhu, *Context-dependent data envelopment analysis - Measuring attractiveness and progress*, The International Journal of Management Science, vol. 31, pp. 397-408, 2003.

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

Analysis and Prediction of crossing effect on inherent deformation during the line Heating Process – Part 2 – multiple crossed heating lines

Adan Vega Saenz

International Maritime University of Panama

emails: avega@umip.ac.pa

Abstract

During the plate forming by line heating, a large amount of heat is applied to the metal surface in order to attain certain curvature. Although this is the best way existing to bend a thick steel plate, the lack of knowledge about the mechanism of forming causes delay and over cost, thus automation is needed. In this paper, the effect of previous heating on inherent deformation by following heating, more specifically, the case of heating lines intersecting (crossing) each other is studied in details. The case of single crossed heating lines was examined in Part 1 of the paper. The second part discusses the case of more than one crossed heating lines. Experiments of multiple heating lines are performed to validate the mathematical model. The influence of multiple crossing in inherent deformation is explained, and new relationship between deformation and crossing effect are presented. Finally, some conclusion of this study are drawn.

Key words: ine heating, crossed heating lines, crossing effect, multiple crossed heating lines, plate forming

1. Introduction

Line heating is one of the most significant plate forming processes used in the shipbuilding industry. However, the line heating process is far from been fully automated causing delays in production, even when some attempts have been

 ISBN: 978-84-608-6082-2

made. The main reason of this is the fact that the relation between applied heat and final plate deformation, the key to automating the process, is too complicated to analyze by using simple mechanical models. Aiming to solve this, many researchers has presented theories to explain both the thermal and mechanical problem (eg. Jang, Seo and Ko (1997), Chang, Liu, and Chang (2005), Ling and Atluri (2006), Osawa, Hashimoto, Sawamura, Kikuchi, Deguchi, and Yamaura (2007), Liu (2006), Liu, Liu, and Hong (2007)). However, most of these researches have been focused on the numerical analysis of single heating lines applied to small plates, which may result in unreliable results. The authors have proposed some improvement in the existing models, considering a series of factors affecting the plate forming such as the geometry of the plate, the cooling condition, the location of the heating, the heat-induced curvature, the residual stresses, the material properties, and the inter-heating temperature (See Vega et all (2009) and Vega et all (2011)). Real plate sizes and heating condition were implemented. Results of numerical analysis and experiment showed better agreement when those factors were considered. More recently, a series of paper considering multiples heating lines, such as the case of overlapped and parallel heating lines were published (Vega et all (2013)).

In Part 1 of this paper (Vega et all (2015)), our discussion focused on the crossing effect produced by two crossed heating lines. However, in plate forming by line heating, multiple heating lines are applied in different directions until the plastic strain necessary to form the plate is attained. On the other hand, when two or more heating lines are applied close to each other, the resulting deformation depends on the separation between heating lines. For this reason, the crossing effect produced by multiple heating lines may differ from that of single heating lines, as it would be explained in details in this paper.

2 Formulation of the 3 thermal elastic plastic model

The same FEA developed in the first part of the paper has been used to study the influence of crossing on inherent deformation of multiples crossed heating lines. Plate geometry as well heating and cooling conditions are similar to those utilized in the first part. More details of the FEA, the heating, and cooling condition, and the material properties are found in the first part of the paper and in Vega et all (2011).

2.1 Validation of the EA

In order to validate our numerical model we realized experiments of multiple line heating. Figure 1 shows the plate model used in the experiment. Where, three heating lines, named as heating A, B and C are applied. The heating sequence is as follows: B $\rightarrow$ C $\rightarrow$ A. During the experiment, before applying heat, the plate was cooled down to room temperature. In this way, the influences of crossed heating lines were evaluated for the same experiment avoiding influences of inter-heating temperature. The heating and cooling conditions, as well the geometry of the heating source, are the same for all the heating lines. In the beginning, small holes were drilled along both sides of each heating line, on the top and bottom surface, as it is seen in the figure. To evaluate the plate deformation, distances between the measuring points are measured before and after each

heating. Both transversal and longitudinal direction are considered. Thus, shrinkage and bending produced by each heating line are subsequently evaluated.
Figure 2 shows a comparison of the distribution of the transverse shrinkage obtained from experiments and that obtained by simulation of a straight heating line. Where, simulation results agree quite well with experimental results. Both, the average deformation and its variation along the heating line are well captured by the numerical model. The cross effect, as well as the influence of the separation between crossed areas on cross effect, are clearly observed in the figure. It is important to note that the complexity of this kind of experiment is high. In addition, it is costly and time-consuming, reducing the possibility of obtaining large number of data.

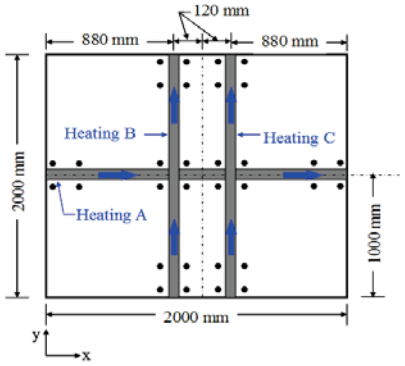


Figure 1. Plate model used to estimate deformations produced by multiple heating lines during experiments of plate forming by line heating

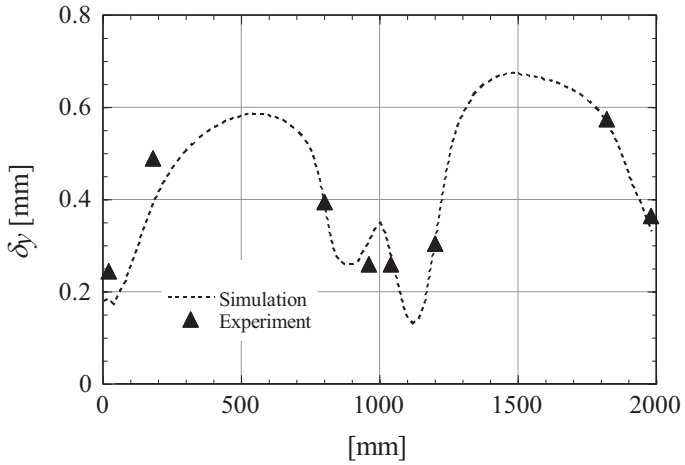
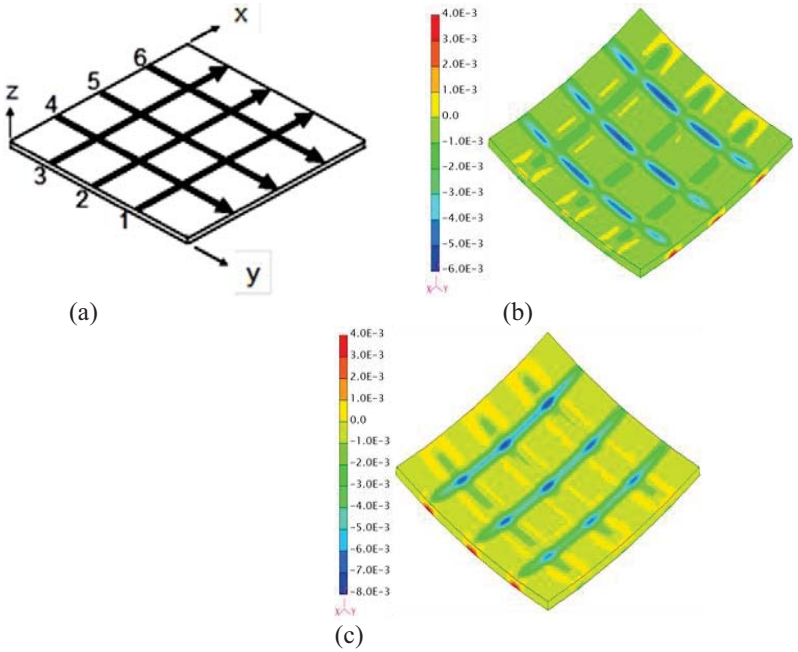


Figure 2 Transverse Shrinkage obtained from experiments and by simulation of multiples heating lines (Plotted along heating A, Figure 1)

3. Inherent deformation due to multiple heating lines

In part one of the paper it was concluded that residual stresses produced by previous heating affect the inherent deformation of subseqent heating. The case of two crossing

heating lines, under different heating conditions, is discussed in detail. However, and as is mentioned above, in plate forming by line heating, several heating lines need to be applied. Thus, it is correct to think about the cross effect for multiples crossing points, or in another worlds, the effect of complex patterns of residual stress in the inherent deformation. Figure 3a shows and example of multiple heating lines. The heating sequence is following the numbers order in the figure. Figures 3b and c show the distribution of plastic strain plotted after the last heating. In all cases, plate was cooled down to roon temperature before apply the next heating. As seen in both figures, the distribution of plastic strain is complex, specially at the crossing areas. Two important points are drawn from these figures: first, inherent deformation of multiple heating line is not simple and second, the cross effect for multiple heating lines is not the same at each crossed area. This study is performed by using a 3000 x 3000 x 40 mm mild steel plate. Noted that this plate is larger than the one use to validate the FEA (2000x2000x40mm). It is important to mentioned that when small plates (smaller than 1000 mm) are used, the inherent deformation is understimated (Vega (2009)).



(a) (b)

(c)

Figure 3 Example of múltiples heating lines (a) Sequence of heating, (b) Plastic strain (ε_x component), and (c) Plastic strain (ε_y component)

4. Residual stress due to multiples heating lines

As shown in previous section, the distribution of plastic strain in a plate where multiples heatng lines are applied is complex, as a result, the inherent deformation and the cross effect may be carefully examined. Let us consider the same plate model shows in Figure 3a, but instead of using the same heating pattern, six parallel heating lines are first applied. The first heating line is applied at 500 mm from plate side edge, and the

following lines are applied parallel, spaced at 400 mm from each other. The distribution of residual stresses produced by each heating line is expected to be similar to that shown by plotting the residual stress distribution of a single heating line. However, when parallel heating lines are applied closely to each other, the existing residual stress is influenced by the thermal cycle of the new heating line and therefore a new pattern of residual stress, different to the one produced by single heating is produced. Similar behavior is observed by application of additional parallel heating lines. Figure 4 shows the distribution of residual stress in and *y* direction obtained after applying the last heating line. It may be noted that at the position of the last heating line, the tensile residual stress is higher (similar to the one obtained from a single heating), while at earlier heating lines, it is smaller. It may also be observed that outside of the heated areas, compressive residual stress in direction and tensile residual stress in *y* direction are larger than in the case of single heating line (See part one of the paper). The variation of residual stress changes with the separation between parallel heating lines as shown in Figure 5 where the case of parallel heating lines spaced 200 mm from each other, applied over the same plate, is shown. Here it can be seen that for smaller separation between parallel heating lines, the larger amount of residual stresses accumulates far from the heated area, while at the heated area residual stresses are smaller. In both cases, 400 and 200 mm of separation between parallel heating lines, the pattern of residual stresses is different so it is expected that the cross effect is also different. In the following section, the cross effect produced by multiples heating line is studied in detail.

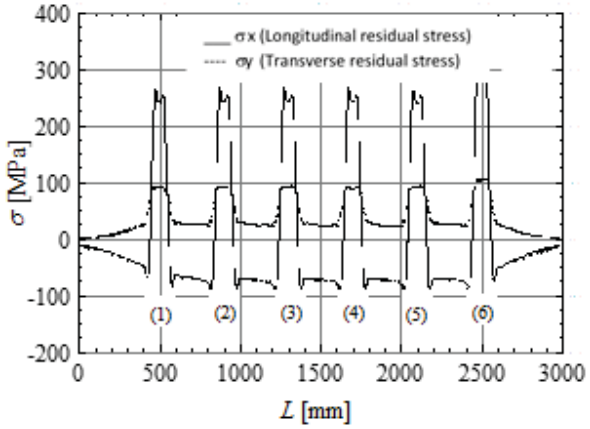


Figure 4 Distribution of residual stresses produced by parallel heating lines spaced 400 mm

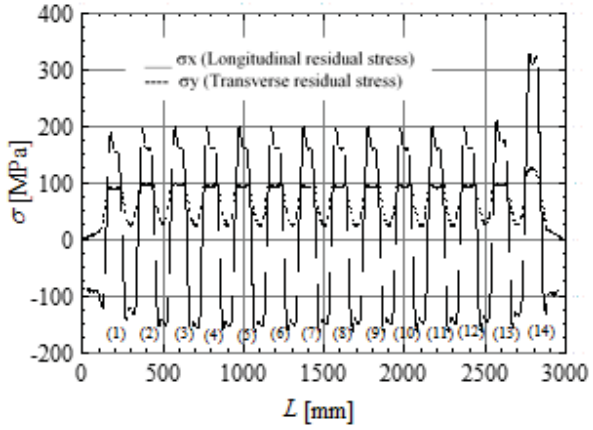


Figure 5 Distribution of residual stress due parallel heating lines spaced 200 mm

5. Crossing effect due to multiple crossing heating lines

As explained in previous section, when multiple heating lines are applied, the residual stress pattern vary depeding on the separation between heating lines. In analyzing the cross effect produced by multiple heating lines it may be necessary to take into account two important aspects as follows: 1) The residual stress at the crossing area will be smaller while that outside of the heating are is larger than that produced by a single heating. 2) The fact that the residual stress outside of the heated area increases with multiples heating lines, means that the inherent deformation produced not only at the heated areas but also far from that, is influenced by residual stresses. In order to introduce these two points in the analysis of inherent deformation we need to separate the crossing effect into two components: that caused by the residual stress in the crossing area (heated area), and that caused by the residual stress existing outside of the heated area, otherwise it will be confused and difficult to estimate.

Figure 6 shows the crossing effect on inherent deformation in case of parallel heating lines spaced 400 mm from each other. The heating condition are the same as previous. Note that the cross effect on each component oh inherent deformation follows similar trend that the pattern of residual stresses showed in Figure 4, been the one produced at the last crossing area the larges. In addition, the four components of the cross effect, at the crossed area, are smaller compared with the cross effect produced by single heating (Figure XXX in part one of the paper), while outside of the crossed area are larger. Similar trend was also noted when performed the same study but changing the separation between heating lines. It is important to remark that in each case of this study, we keep both the separation and the heating condition constants. If any of these vary, the cross effect vary. Considering all possible combinations of heating condition and separation between heating line would require additional extra work that is out of the scope of this paper.

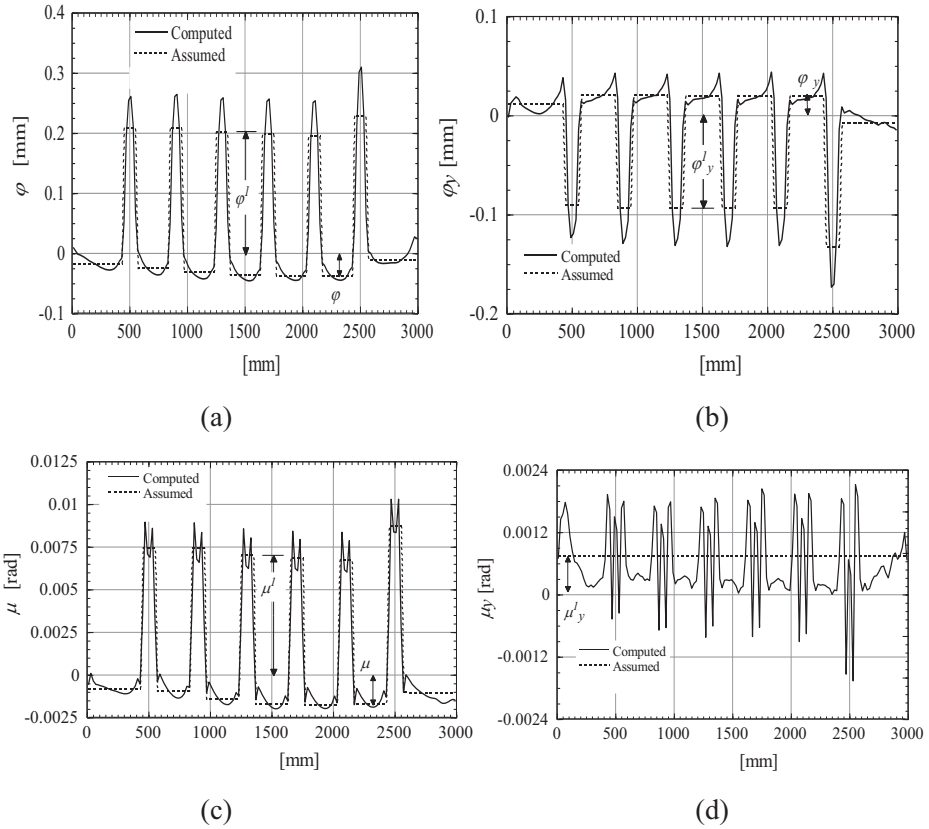


Figure 6 Distribution of crossing effect on inherent deformation for heating lines spaced 400 mm (a) Longitudinal shrinkage, (b) Transverse shrinkage, (c) Longitudinal bending and (d) Transverse bending

In case of multiple heating lines the cross effect shows a complex distribution as seen in Figure 6. In addition, it become difficult to estimate the cross effect for each crossing area separately. To overcome this, we introduce the concept of total inherent deformation which is given by the integration of the plastic strain over the whole plate as is given by Equation 1 to 4.

$$\delta^t = \iiint \varepsilon^* d \; dydz / h \qquad (\text{mm}^2) \qquad (1)$$

$$\delta_y^t = \iiint \varepsilon_y^* d \; dydz / h \qquad (\text{mm}^2) \qquad (2)$$

$$\theta^t = \iiint \varepsilon^* (z - h/2) / (\text{h}^3/12) d \; dydz \quad (\text{radians . mm}) \qquad (3)$$

$$\theta_y^t = \iiint \varepsilon_y^* (z - h/2) / (h^3/12) d \; dydz \quad (\text{radians . mm}) \qquad (4)$$

By using these equations, the total cross effect for multiples heating lines can be determined assuming that it distribute uniform along each heating line, as follows:

$$\varphi_{\text{x}} = \delta^t - \delta^{t1} - \delta^{t2} - .....\delta^{tn} \qquad (\text{mm}^2) \qquad (5)$$

$$\varphi^t_y = \delta^t_y - \delta^{t1}_y - \delta^{t2}_y - .....\delta^{tn}_y \quad (\text{mm}^2) \tag{6}$$

$$\mu^t_x = \theta^t - \theta^{t1} - \theta^{t2} - .....\theta^{tn} \quad (\text{Radians . mm}) \tag{7}$$

$$\mu^t_y = \theta^t_y - \theta^{t1}_y - \theta^{t2}_y - .....\theta^{tn}_y \quad (\text{Radians . mm}) \tag{8}$$

Once we know the total cross effect, it will be necessary to separate it into the tow components mentioned before. The resulting cross effect is an approximation of the real one, however if all the parameter are considered, accurately values are obtained by this way.

6. Relationship between the cross effect and the separation of parallel heating lines

The same heating pattern used during experiments of line heating (Figure 1) was later used to simulate the variation of cross effect with separation between parallel heating lines. The heating and cooling conditions are the same for all the cases thus, the cross effect is mainly dependent on the pattern of residual stresses of each particular case. Figure 7 shows the results. Herein, we named the separation between parallel heating lines with the letter *c* Note that in these figures, the two components of crosse effect (that produced at the crossing area (solid lines) and that produced outside the crossing area (in dashed lines) are presented. Values on the figures correspond to the total cross effect, thus may not be compared with previous figures.

Conclusions

The primary objective of this paper was to develop a new method that can improve the precision in predicting plate deformation during the plate forming by line heating. In order to accomplish that objective, a FEA has been performed to study the influence of multiple crossed heating lines on inherent deformation during the plate forming by line heating. From the results of this study the following conclusions are drawn:

1. Experiments of multiple heating lines were performed using the induction heating system IHI-α. The comparison between simulation and experiments shows good agreement. In addition, the cross effect as well as the influence of separation between crossed areas on cross effect are well captured in the experiments. Thus, demonstrate our suggestion, that it is necessary to consider the cross and parallel effect in the analysis of multiple heating lines is correct.
2. The inherent deformation produced by multiple crossed heating lines applied close to each other is influenced by residual stresses produced by previous heating lines.
3. The concept of crossing effect proves to be useful when considering the effect of multiple crossed heating lines on inherent deformation. An inherent deformation database of crossing effect is proposed as an alternative for prediction plate deformation.

Acknowledgments

Financial support from The Secretary of Science and Technology (SENACYT) of The Government of The Republic of Panama and by Class IBS is gratefully acknowledged.

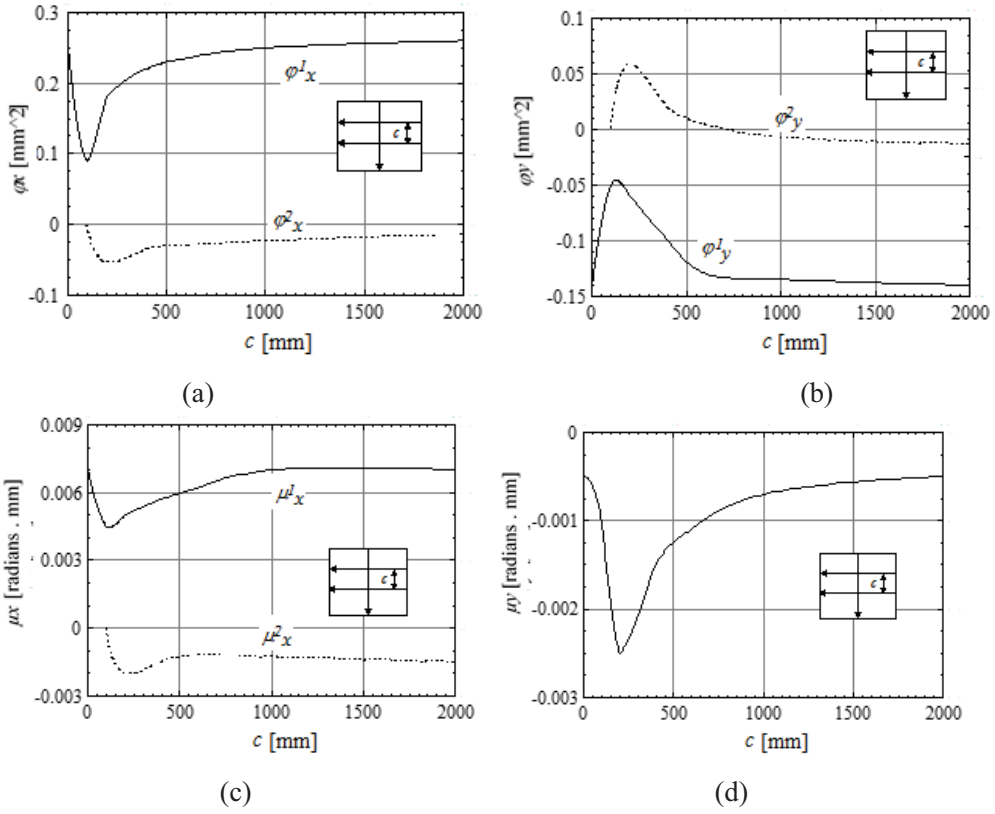


Figure 7 Relation between crossing effect on inherent deformation and separation between parallel heating lines for both, at the cross area (solid lines) and at outside of the cross area (dashed lines) (a) Longitudinal shrinkage, (b) Transverse shrinkage, (c) Longitudinal bending and (d) Transverse bending

2. References

[1] Chang, C.W.; Liu, C.S. and Chang, J.R. (2005): A Group Preserving Scheme for Inverse Heat Conduction Problems. *C : Computer odeling in ngineering ciences*, 10, 1, pp.13-38.

[2] Jang, C. D.; Seo, S.; Ko. D. E. (1997): A study on the prediction of deformations of plates due to line heating using a simplified thermal elasto-plastic analysis. *ournal of hip Production,* 13(1):22-27.

[3] Liu, C.S. (2006): An Efficient Simultaneous Estimation of Temperature-Dependent Thermophysical Properties. *C : Computer odeling in ngineering ciences*, 14, 2, pp.77-90.

[4] Liu, C.S.; Liu L.W. and Hong, H.K. (2007): Highly Accurate Computation of Spatial-Dependent Heat Conductivity and Heat Capacity in Inverse

Thermal Problem. *C : Computer odeling in ngineering ciences*, 17, 1, pp.1-18.
[5] Ling, X. and Atluri, S.N. (2006): Stability Analysis for Inverse Heat Conduction Problems. *C : Computer odeling in ngineering ciences*, 13, 3, pp.219-228.
[6] Osawa, N.; Hashimoto, K.; Sawamura, J.; Kikuchi, J.; Deguchi, Y. and Yamaura, T. (2007): Development of Heat Input Estimation Technique for Simulation of Shell Forming by Line-Heating. *C : Computer odeling in ngineering ciences*, 20, 1, pp.45-53.
[7] Vega, A., Rashed, S., Tango, Y., Ishiyama, M., Murakawa, H.: Analysis and prediction of multi-heating lines effect on plate forming by line heating. *C ournal: Computer odeling in ngineering ciences, C* , Vol. 28, No. 1, pp. 1-14, 2008
[8] Vega, A., Escobar, A., Fong, A., Ma N., and Murakawa, H. Analysis and Prediction of Overlapped effect on Inherent Deformation during the Line Heating Process. Journal of Computer Modeling in Engineering & Sciences, CMES. March, 2013
[9] Vega, A., Camaño, A., Fong, A., Ma N., and Murakawa, H. Analysis and Prediction of Parallel effect on Inherent Deformation during the Line Heating Process. Journal of Computer Modeling in Engineering & Sciences, CMES. February, 2013
[10] Vega, A.; Tango, Y.; Ishiyama, M.; Rashed, S.; Murakawa, H.: Influential Factors Affecting Inherent Deformation during Plate Forming by Line Heating (Report 5) – The Effect of water cooling. International Journal of Offshore and Polar Engineers (IJOPE), June, 2011
[11] Vega, A.; Tango, Y.; Ishiyama, M.; Rashed, S.; Murakawa, H.: Influential Factors Affecting Inherent Deformation during Plate Forming by Line Heating (Report 4) – The Effect of material properties. International Journal of Offshore and Polar Engineers (IJOPE), March, 2011
[12] Vega, A., Osawa, N., Rashed, S., and Murakawa, H. Analysis and Prediction of Edge Effect on Inherent Deformation of Thick Plates Formed by Line Heating. CMES Journal. March 2011.
[13] Vega, A. (2009); Development of Inherent Deformation Database for Automatic Forming of Thick Steel Plates by Line Heating Considering Complex Heating Patterns. *Doctoral hesis*. Osaka University, Japan, 2009.

 ISBN: 978-84-608-6082-2

Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016
4-8 July, 2016

A New Method for Calculation of Density of Compressible Flow

Tiejin Wang
China Academy of Aerospace Aerodynamics
emails: tiej701@163.com

Abstract

According to the equations of the propagation path of the light passing through the non-uniform field of density, a new method for calculating the density of the compressible flows is presented in the paper. The principle of the new method is mainly introduced in this abstract.

Key words: density, calculation method, compressible flow, experiment method

1. Introduction

The density calculation of a compressible flow is very important for the aero-optics, because the theoretical analysis and computation of the optic transmission effects are all based on the data of the density field. Usually, the compressible flow is described by the nonlinear N-S equations, due to its complexity, at present there is still no theoretical solution for the general equations and only some solutions exists for its simplified forms. In most cases, the density is now calculated using the CFD method, in which the N-S equations is treated through discrete and numeral methods, but the calculated results of the density field must be testified by experimental results.

Up to now, a few of density measurement methods can be used in the compressible flows, for example, the interference method for the two-dimensional continuous density field, the electron beam method for the low-pressure density field and so on. Most of the density fields of the compressible flows can't be measured.

In the compressible flows, the changes of gas density will lead to the changes of its refractive index, so that when a beam of light passes through the gas, the light propagation path will be changed. According to the equations of the propagation path of the light passing through the non-uniform field of density, a new method for calculating the density of the compressible flows is presented in the paper. Due to the density term included in the path equations of light, when the density

 ISBN: 978-84-608-6082-2

distribution is known, the optical path can be obtained by the integrating the equations. In the new method, the optical path is known, so the density of flow can be calculated through the optical path equations. The known optical paths are obtained by experiments.

2. Principle of new method

The new method is a combination method of theory and experiment, where two parts are included: one is the path equations of light, the basis of the new method; another is the experimental method to obtain the optical path, the key of the method.

(1) Path equations of light

The propagation of the light through a continuously changed field of refractive index can be analyzed using the Fermat Principle[1, 2], namely that the differential of the optical path length must be zero, then

$$\partial \int n(x, y, z)ds = 0 \qquad (1)$$

Where the s is the length of arc, n is the gas refractive index. This variation equation can be written as the following three Euler-Lagrangian equations,

$$\frac{d}{ds}\left(\frac{\partial F}{\partial x'}\right) = \frac{\partial F}{\partial x}$$
$$\frac{d}{ds}\left(\frac{\partial F}{\partial y'}\right) = \frac{\partial F}{\partial y}$$
$$\frac{d}{ds}\left(\frac{\partial F}{\partial z'}\right) = \frac{\partial F}{\partial z} \qquad (2)$$

Where，$F=F(x, y, z, x', y', z', s) = n(x, y, z)(x'^2+y'^2+z'^2)^{1/2}$

$ds^2=dx^2+dy^2+dz^2$, $(x'^2+y'^2+z'^2)^{1/2}=1$,

$$x' = \frac{dx}{ds}, \quad y' = \frac{dy}{ds}, \quad z' = \frac{dz}{ds}$$

Then the above Euler-Lagrangian equations can be written as

$$\frac{d}{ds}\left(n \cdot \frac{ds}{dx}\right) = \frac{\partial n}{\partial x}$$

 ISBN: 978-84-608-6082-2

$$\frac{d}{ds}\left(n\cdot\frac{ds}{dy}\right)=\frac{\partial n}{\partial y}$$
$$\frac{d}{ds}\left(n\cdot\frac{ds}{dz}\right)=\frac{\partial n}{\partial z} \quad (3)$$

When the light enters the density field parallel to the z direction, the parameter s can be eliminated, and x and y can be treated as the function of z, then the following differential equation group is obtained,

$$\frac{d^2x}{dz^2}=\left[1+\left(\frac{dx}{dz}\right)^2+\left(\frac{dy}{dz}\right)^2\right]\left[\frac{1}{n}\frac{\partial n}{\partial x}-\frac{dx}{dz}\cdot\frac{1}{n}\frac{\partial n}{\partial z}\right]$$
$$\frac{d^2y}{dz^2}=\left[1+\left(\frac{dx}{dz}\right)^2+\left(\frac{dy}{dz}\right)^2\right]\left[\frac{1}{n}\frac{\partial n}{\partial y}-\frac{dy}{dz}\cdot\frac{1}{n}\frac{\partial n}{\partial z}\right] \quad (4)$$

Generally, the gas density, ρ, and the gas refractive index, n, can be related by the Gladstone-Dale formula as

$$n=1+K_{GD}\rho \quad (5)$$

where the constant K_{GD} shows a kind of gas character. When the formula (5) is substituted into equation (4), then

$$\frac{d^2x}{dz^2}=K_{GD}\left[1+\left(\frac{dx}{dz}\right)^2+\left(\frac{dy}{dz}\right)^2\right]\left[\frac{1}{1+K_{GD}\rho}\frac{\partial\rho}{\partial x}-\frac{dx}{dz}\cdot\frac{1}{1+K_{GD}\rho}\frac{\partial\rho}{\partial z}\right]$$
$$\frac{d^2x}{dz^2}=K_{GD}\left[1+\left(\frac{dx}{dz}\right)^2+\left(\frac{dy}{dz}\right)^2\right]\left[\frac{1}{1+K_{GD}\rho}\frac{\partial\rho}{\partial y}-\frac{dy}{dz}\cdot\frac{1}{1+K_{GD}\rho}\frac{\partial\rho}{\partial z}\right] \quad (6)$$

The path of light passing through the non-uniformity density field can be described by the integration of the above equation group.

(2) Experimental methods for optical path

The experimental method for the optical path can be described as the following: a beam of light passing through the density field of a compressible flow to be measured becomes curved due to the change of density, and the optical path in the non-uniformity density field comes into being; using the visualization methods, the optical path can be seen; when the light beam id thin enough, the optical path can be looked as an spatial curve; three sets of cameras in three different positions are used to synchronously recorded the photos of the optical path; using the

 ISBN: 978-84-608-6082-2

photograph processing and reconstruction technologies, the spatial coordinates of the optical path are obtained and the optical path is gotten by experiment.

Beside the above introduction of the principle of the new method are also presented the feasibility of the new method, the application of the experimental results and the error range of the new method in the paper.

3. References

[1] M. BORN, E. WOLF, OPTICAL PRINCIPLE, TRANSLATED BY LETIAN HUANG, SCIENCE AND TECHNOLOGY PRESS, BEIJING, 1978 (IN CHINESE)

[2] GUICHUN LI, OPTICAL INSTRUMENTATION FOR WIND TUNNEL TESTING, NATIONAL DEFENSE INDUSTRY PRESS, MAY 2008 (IN CHINESE)

 ISBN: 978-84-608-6082-2

Late Contributions

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2016
4–8 July, 2016.

An almost fully parallel algorithm for computing the component tree of a binary digital image based on HSF

F. Diaz-del-Rio[1], P. Real[1] and Darian Onchis[2]

[1] *HTS. Informatics Engineering, Universidad de Sevilla (Spain)*

[2] *Faculty of Mathematics, University of Vienna (Austria)*

emails: fdiaz@us.es, real@us.es, darian.onchis@univie.ac.at

Abstract

Based on the works [1] and [2], we focus here on the problem of parallel computing the component topological tree ([3]) of a binary digital image I of dimension nxm. We use 4-adjacency for black pixels and 8-adjacency for white pixels. Using as many elementary unit processing as pixels the digital image has and taking as mathematical model of I the graph-based structures of [1, 2] called homological spanning forest (HSF, for short), we design and implement an almost fully parallel algorithm for computing the topological (component) rooted tree of I.

Key words: 2D digital image, component tree, parallelism

1 Introduction

A parallel computational framework for topological computation within 2D digital image context is developed in [1, 2]. The two main steps of any topological processing in such framework are in order:

- **Input data and extraction of the ROI.** In order to avoid segmentation and noise issues which are ubiquitous mathematically ill-posed problems in the area of Digital Imagery, the input data are 2-dimensional integer-valued matrices associated to a binary or gray-level presegmented 2D digital image I. In a binary digital image, the ROI is specified by the set of black pixels.

 ISBN: 978-84-608-6082-2

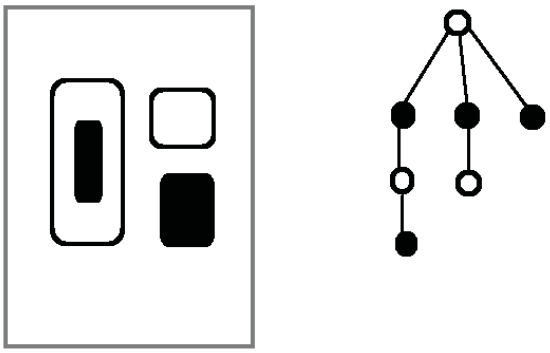


Figure 1: Example of topological connected tree of a digital image

- **Elementary sequential or parallel step; generation of homological spanning forests (HSFs, for short) of the ambiance discrete space**. After embedding the pixel-based digital image within a much more bigger regular ambiance sub-pixel scenario, the computational strategy is to generate in a parallel way a kind of dense graph skeleton of I, called HSF. More precisely, we are interested in a HSF of the digital image, such that its restriction to the ROI D has a minimal number of 4-connected components. For achieving this, we use techniques of parallel "transport" of interactions between pixels through the graph-based scaffolding of the HSFs.

Our interest here is to design and to implement a parallel algorithm based on the previous approach for computing the topological component tree of binary digital image. If we determine that the set of black pixels is the ROI D of I, we employ 4-adjacency between pixels of D and 8-adjacency for pixels of $I \setminus D$.

2 Component tree of a 2D binary image

The component tree of a 2D binary digital image I (being D the set of black pixels) is the rooted tree having as alternate level nodes the 4-connected component of black pixels and the 8-connected component of white pixels, and as edges the relationships between them of the kind "to be sourrounded by". An example of component tree of a binary image is shown in Figure 1. The root of the component tree is always represented by the dummy white region in which it is embedded the whole binary image. In Figure 2, the corresponding black

4-connected components (green squares, sinks) and 8-connected components (red squares, sources) of a face-like binary image are identified. This topological recognition is based on the parallel optimization of HSF structures given in [2], in which resulting HSFs have always the minimal number of black 4-connected components (see Figure 3).

Finally, the component tree of the image is deduced from this final HSF, in such a way that the critical sinks are connected through the HSF to the resulting sources, Due to the fact that the numbers of these sinks and sources are very small with regards to the total size mn of the image, this process can be efficiently and fastly done. Finally, the results given by the algorithm for generating the component tree in this concrete case are the following:

$$\text{Number of sinks}= 6; \text{number of sources}= 4.$$

Table of sinks:

92	61
165	33
194	163
200	33
226	33
258	17

Table of sources:

33	258
61	226
163	226

Determination of the Component Tree:

	33	61	163
92	0	−1	0
165	−1	0	0
194	0	0	−1
200	−1	0	0
226	−1	1	1
258	1	0	0

3 Implementation.

In order to check the speed and scalability of the algorithm, a non-fully functional but complete implementation has been done in C++ using OpenMP directives.

The fundamental topological tools have been transformed so as to promote an efficient parallel implementation in any parallel-oriented architecture (GPUs, multithreaded

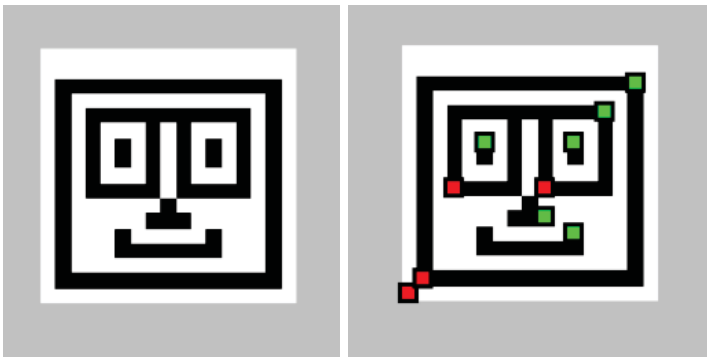


Figure 2: Left: Original face-like image.Only 4-adjacency is valid for Foreground (black) pixels See that the nose is not connected with the glasses. Right: Sinks (red squares) and sources (red squares) of the face. An additional source in the most left-bottom corner has been added to represent a virtual foreground source for the whole image.

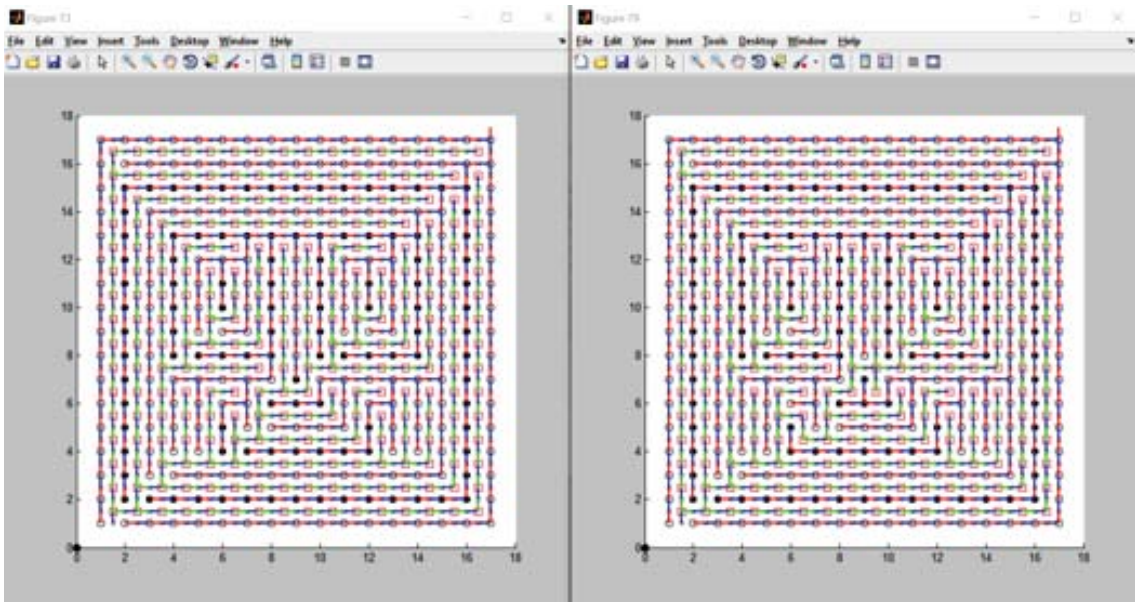


Figure 3: Left:Initial HSF of the binary image; Right: Final HSF having the minimal number of black 4-connected components. The component tree can be automatically extracted from this structure.

computers, SIMD kernels and so on), by taking advantage of the inherent massive data parallelism that any image processing has.

The time complexity of this algorithm is of the order of the logarithm of the sum of the width and the height of the image. Only a linear term appears on the last sequential steps (including the determination of the edges of the rooted component tree). Moreover, the algorithm is almost fully parallel, so it is expected to scale well for any parallel architecture (GPUs, SIMD kernels, multithreaded, etc.).

References

[1] H. Molina-Abril, P. Real, *Homological spanning forest framework for 2d image analysis*, Annals of Mathematics and Artificial Intelligence **64** (2012) 385409.

[2] F. Diaz-del-Rio, P. Real, D. Onchis, *A parallel Homological Spanning Forest framework for 2D topological image analysis*, Accepted in Pattern Recognition Letters(2016)

[3] R. Cucchiara, C. Grana, A. Prati, S. Seidenari, G. Pellacani , *Building the topological tree by recursive FCM color clustering*, In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 1, pp. 759-762). IEEE.